Part I

Current nitrogen fertilization recommendations for wheat include applications of fertilizer at specified stages of plant growth. An experiment was conducted to evaluate the effect of two fertilization timing schedules (Schedule 1 and Schedule 2) on plant nitrogen content measured by the stem tissue nitrate amount. The experiment was carried out in an irrigated field divided into four blocks such that the plots within each block were in the same part of the water gradient. In each block, one plot was randomly selected and assigned to Schedule 1. The other plot in the same block was assigned to Schedule 2. The same fertilizer amount (at a low level) was applied to all plots. The observed nitrate nitrogen content (ppm × 10^{-2}) from each plot is shown in Table 1.

Table 1: Observed nitrate nitrogen content (ppm × 10^{-2})

Block	Schedule 1	Schedule 2
1	34.98	37.18
2	41.22	45.85
3	36.94	40.23
4	39.97	39.20

1. Use a Wilcoxon signed rank test or a sign test to assess if the fertilization timing schedule has an effect on the nitrate nitrogen content. Specify the following in your answer:
(i) the test that you choose to apply,
(ii) the exact p-value, and
(iii) your conclusion in the context of the study.

Part II

The data from **Part I** are a subset of a dataset from a larger study. In the larger study, the experiment was conducted to evaluate the effects of the two fertilization timing schedules (Schedule 1 and Schedule 2) and three fertilizer amounts (Low, Medium, and High) on stem tissue nitrate amounts of wheat plants. The experiment was carried out in a field divided into four blocks. Each of the six plots within each block was randomly assigned to one of the six combinations of schedule and fertilizer amount. The observed nitrate nitrogen content (ppm × 10^{-2}) from each plot is given in Table 2.

Table 2: Observed nitrate nitrogen content (ppm × 10^{-2})

Block	Low Schedule 1	Low Schedule 2	Medium Schedule 1	Medium Schedule 2	High Schedule 1	High Schedule 2
1	34.98	37.18	37.99	34.89	40.89	42.07
2	41.22	45.85	41.99	50.15	46.69	49.42
3	36.94	40.23	37.61	44.57	46.65	52.68
4	39.97	39.20	40.45	43.29	41.90	42.91

The column headings of "Low, Medium, and High" correspond to the levels of fertilizer amount.

Index the blocks by $i = 1, 2, \ldots, 4$, the schedules by $j = 1, 2$, and the fertilizer amounts by $k = 1, 2, 3$ for Low, Medium, and High, respectively. Let Y_{ijk} be the observed nitrate nitrogen content for block i, schedule j, and fertilizer amount k. Suppose

$$Y_{ijk} = \mu + \beta_i + \alpha_j + \gamma_k + (\alpha\gamma)_{jk} + \varepsilon_{ijk} \qquad \text{Model (1)}$$

where μ, α_j, γ_k, and $(\alpha\gamma)_{jk,}$ are unknown real-valued parameters, β_i are i.i.d. $N(0, \sigma_\beta^2)$ random variables, ε_{ijk} are i.i.d. $N(0, \sigma^2)$ random variables, and the β_i are independent of the ε_{ijk} for all ijk. The data step of SAS code is given below and partial output from SAS for fitting Model (1) to the data of Table 2 is on pages **4-5**.

```
data nitrate;
  input block schedule fertilizer Y;
  datalines;
1 1 1 34.98
1 1 2 37.99
1 1 3 40.89
1 2 1 37.18
1 2 2 34.89
1 2 3 42.07
2 1 1 41.22
2 1 2 41.99
2 1 3 46.69
2 2 1 45.85
2 2 2 50.15
2 2 3 49.42
3 1 1 36.94
3 1 2 37.61
3 1 3 46.65
3 2 1 40.23
3 2 2 44.57
3 2 3 52.68
4 1 1 39.97
4 1 2 40.45
4 1 3 41.90
4 2 1 39.20
4 2 2 43.29
4 2 3 42.91
  run;
```

2. Provide the "Source" and corresponding "Degrees of Freedom" columns for the ANOVA table on page 4 for the fit of Model (1) to the data in Table 2.
3. What are the default baseline constraints that SAS sets on the model parameters α_j , γ_k , and $(\alpha\gamma)_{jk}$?
4. Give the mean parameter vector corresponding to the constraints you identified in problem **3** and give the full-rank design matrix **X** corresponding to your mean parameter vector for the first six observations listed above in SAS code.
5. With respect to the mean nitrate nitrogen content for the six combinations of schedule and fertilizer amount, how should the following parameters be interpreted?

   (i) α_1

   (ii) $(\alpha\gamma)_{12}$

6. What is the null hypothesis for testing an interaction between Schedule and Fertilizer Amount? State the null hypothesis in terms of the parameters of Model (1) with restrictions defined in problem **3**.
7. Conduct a test for the interaction between Schedule and Fertilizer Amount. Compute the value of the test statistic and give its degrees of freedom. *Note that you need not simplify any numerical expression after you plug in all numbers when you are asked to compute a value. This note also applies to all remaining problems (**Problems 7, 9, 12-15** in Part II).*
8. The investigator for this study wants to know whether there is any difference in the mean nitrate nitrogen content between the two schedules when fertilizer amount is low (the research question addressed in **Part I**). Is this difference a simple effect, a main effect, or an interaction effect?
9. Conduct a test to answer the question of whether there is any difference in the mean nitrate nitrogen content between the two schedules when fertilizer amount is low. Compute the value of the test statistic and give its degrees of freedom.
10. Based on the residual plots on page 5, do you have any concerns about the appropriateness of the assumptions for Model (1) for this dataset? Explain.
11. To answer the question whether there is any difference in the mean responses between the two schedules when fertilizer amount is low, do you choose to report results from problem **1** or problem **9**? Justify your answer.
12. Conduct a test for the main effect of fertilizer using Model (1). Give your null hypothesis, the value of your test statistic, and the degrees of freedom for the corresponding F distribution.
13. What is your estimated value of $Var(Y_{ijk})$ using Model (1)?
14. Estimate the correlation between Y_{ijk} and Y_{ijl}, where $k \neq l$, under Model (1).
15. Estimate the correlation between Y_{ijk} and Y_{mjk}, where $i \neq m$, under Model (1).
16. Suppose that we have two plots for each combination of Schedule and Fertilizer Amount in each block, and hence a total of 48 observations. Give the "Source" and corresponding "Degrees of Freedom" columns of the ANOVA table for the fit of a model including an interaction term between block and treatment.

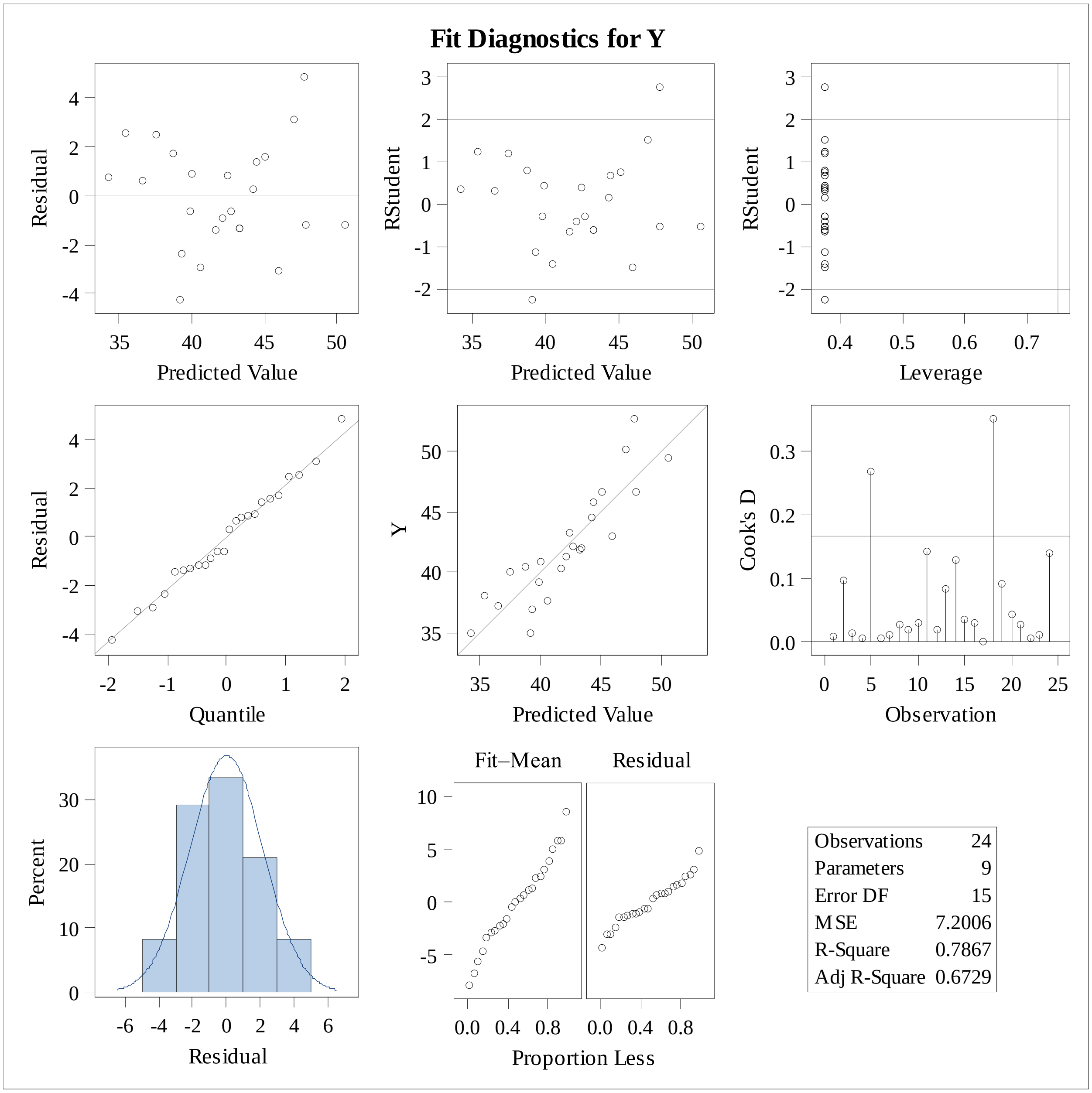
Partial SAS output for fitting Model (1) in Part II

Model Information	
Data Set	WORK.NITRATE
Dependent Variable	Y
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information		
Class	**Levels**	**Values**
block	4	1 2 3 4
schedule	2	1 2
fertilizer	3	H L M

Type 3 Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**Expected Mean Square**
schedule			51.51	Var(Residual) + Q(schedule,schedule*fertilizer)
fertilizer				Var(Residual) + Q(fertilizer,schedule*fertilizer)
schedule*fertilizer			1.00	Var(Residual) + Q(schedule*fertilizer)
block			65.67	Var(Residual) + 6 Var(block)
Residual			7.20	Var(Residual)

Partial SAS output for fitting Model (1) in Part II



Fit Diagnostics for Y

```
Observations        24
Parameters           9
Error DF            15
MSE             7.2006
R-Square        0.7867
Adj R-Square    0.6729
```

Part III

A plant biologist working on Arabidopsis (a model plant studied in the labs) conducted an experiment in a phenotyper where images of plants were taken automatically based on a specified protocol. Information such as height and leaf area were obtained from image analysis. In addition, concentrations of some soil nutrients were measured. One goal of this study is to use the information extracted from image data and soil nutrient levels to predict the biomass (fresh weight) toward the end of growing period. Two watering conditions (well-watered and water-stressed) were used in this experiment. For each condition, observations were collected from 5 plants, each grown in an individual pot. Images were taken in the middle of growing period, and soil nutrients were measured at the same time. Then, the fresh weight was measured at the end of growing period for each plant.

Now, let Y_k be the observed plant weight (in grams) for the k-th pot for $k=1, \ldots, 10$. Let

$$x_{1k} = \begin{cases} 1 & \text{if water condition is well-watered for the } k\text{-th pot} \\ 0 & \text{if water condition is water-stressed for the } k\text{-th pot} \end{cases}$$

x_{2k} = plant height (in cm) measured from the image analysis for the k-th pot

x_{3k} = leaf area (in cm^2) measured from the image analysis for the k-th pot

x_{4k} = soil nitrate concentration (in ppm) for the k-th pot

x_{5k} = soil potassium concentration (in ppm) for the k-th pot

The following regression models were fit to the dataset:

$$Y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \varepsilon_k \qquad \text{Model (2)}$$

$$Y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k \qquad \text{Model (3)}$$

$$Y_k = \beta_0 + \beta_1 x_{1k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \varepsilon_k \qquad \text{Model (4)}$$

For each regression model, the regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4,$ and β_5 are unknown real-valued parameters, and ε_k are i.i.d. $N(0, \sigma^2)$ random variables.

Partial output from SAS for fitting Models 2-4 is on page **7-9**.

17. Interpret the estimated value of the parameter β_1 for model (2) in the context of this study.

18. Based on the diagnostic plots for model (2) on page 8, discuss the appropriateness of Model (2) for these data.

19. Test for the significance of the effects of soil nutrients (both nitrate and potassium) on fresh weight. Specify your null and alternative hypotheses, calculate the value of your test statistic (*you need not simplify any numerical expression after you plug in all numbers*), and give the degrees of freedom.

20. Among Models 2-4, which one do you suggest using for this data analysis? Specify your model selection criterion (criteria), how to evaluate it (them), and how you make your decision.

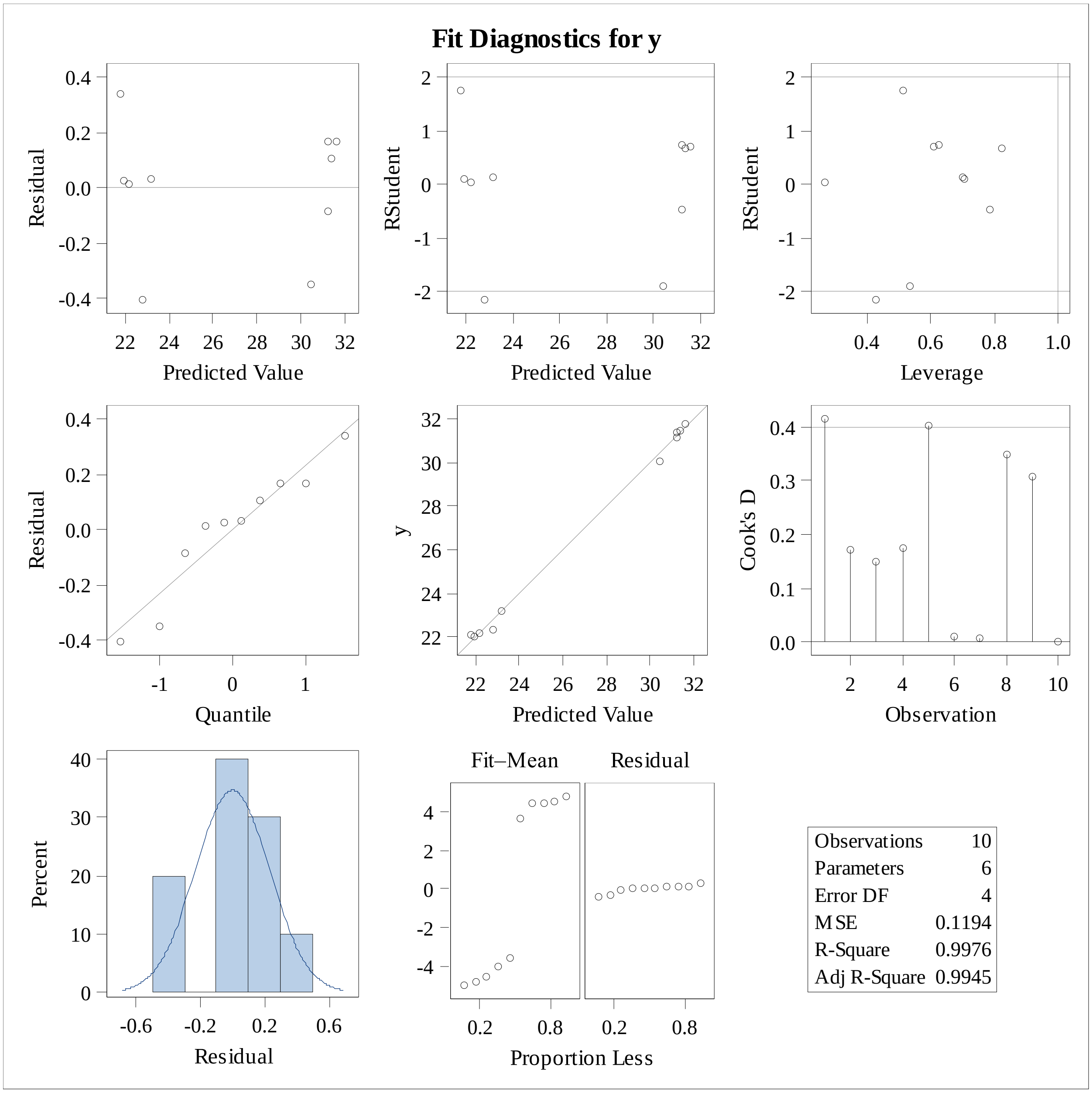
21. After the biologist looked at the regression Models (2-4), he thought an important part is missing. He thinks that the relationship between soil nutrients and fresh weight is likely dependent on the water conditions. Suggest another model to be fit to this dataset in light of his information.

Partial SAS output for fitting Model (2) in Part III

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		194.95			
Error		0.48			
Corrected Total					

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	9.25074	3.48833	2.65	0.0569
x1	1	9.20290	0.29067	31.66	<.0001
x2	1	0.61035	0.25566	2.39	0.0754
x3	1	0.48179	0.56172	0.86	0.4394
x4	1	0.25515	0.07371	3.46	0.0258
x5	1	0.00517	0.00985	0.52	0.6276

Partial SAS output for fitting Model (2) in Part III



Observations	10
Parameters	6
Error DF	4
MSE	0.1194
R-Square	0.9976
Adj R-Square	0.9945

Partial SAS output for fitting Model (3) in Part III

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		193.32			
Error		2.10			
Corrected Total					

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20.89361	2.24532	9.31	<.0001
x1	1	8.68106	0.42365	20.49	<.0001
x2	1	0.45315	0.37583	1.21	0.2733
x3	1	-0.36012	0.87635	-0.41	0.6954

Partial SAS output for fitting Model (4) in Part III

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model		194.17			
Error		1.25			
Corrected Total					

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.22292	3.60163	3.67	0.0104
x1	1	9.22810	0.38325	24.08	<.0001
x4	1	0.17154	0.08288	2.07	0.0839
x5	1	0.01615	0.01155	1.40	0.2115

Part I

Researchers conducted an experiment to learn about the effects of two feed types (1 and 2) on weight gain in pigs. Ten pens, each containing four pigs, were used for the experiment. Each pen contained a single container, called a feeder, in which researchers placed feed for distribution to the four pigs in a pen. A completely randomized design was used to assign feed type 1 to five pens and feed type 2 to five pens. Feed of the assigned type was placed in each pen's feeder each day for a three-week period. Each pig was weighed at the beginning and end of the three-week period. Let y_{ijk} be the weight gained by kth pig in the jth pen treated with feed type i ($i = 1, 2; j = 1, \ldots, 5; k = 1, \ldots, 4$). Assume the model

$$y_{ijk} = \mu + \phi_i + p_{ij} + e_{ijk}, \tag{1}$$

where μ, ϕ_1, and ϕ_2 are unknown parameters, $p_{ij} \sim N(0, \sigma_p^2)$ for some unknown variance parameter σ_p^2, $e_{ijk} \sim N(0, \sigma_e^2)$ for some unknown variance parameter σ_e^2, and all p_{ij} and e_{ijk} terms are independent.

The weight gain data were stored in a vector `y` in `R`, along with information about the feed treatment and pen provided in factors `feed` and `pen`. The code and output on page 2 contains information useful for answering the problems **1** through **4** below. Note that the end of page 2 provides quantiles of t distributions.

1. Model (1) may be written in the form $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}$, where $\boldsymbol{y}$ is the vector of y_{ijk} values (ordered as in the `R` code), $\boldsymbol{\beta} = (\mu, \phi_1, \phi_2)'$, $\boldsymbol{u} = (p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{21}, p_{22}, p_{23}, p_{24}, p_{25})'$, and $\boldsymbol{e}$ is the vector of e_{ijk} values (ordered to match the order of $\boldsymbol{y}$). Using Kronecker product notation, provide expressions for $\boldsymbol{X}$ and $\boldsymbol{Z}$.

2. Provide the value of an unbiased estimator for σ_p^2.

3. Determine the value of the F statistic you would use to test $H_0 : \phi_1 = \phi_2$.

4. Find a 95% confidence interval for $\phi_1 - \phi_2$.

R Code and Output for Part I

```
> length(y)
[1] 40

> y[c(1:4, 37:40)]
[1] 30.6 23.6 30.6 29.6 29.6 28.0 32.5 29.1

> feed
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2
[32] 2 2 2 2 2 2 2 2 2
Levels: 1 2

> pen
 [1] 1  1  1  1  2  2  2  2  3  3  3  3  4  4  4  4  5  5  5  5 
[21] 6  6  6  6  7  7  7  7  8  8  8  8  9  9  9  9  10 10 10 10
Levels: 1 2 3 4 5 6 7 8 9 10

> anova(lm(y ~ feed + pen))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value    Pr(>F)    
feed       1 119.02 119.025 11.8894 0.0016944 **
pen        8 418.02  52.252  5.2195 0.0003843 ***
Residuals 30 300.33  10.011

> mean(y[feed == 1])
[1] 31.69

> mean(y[feed == 2])
[1] 35.14

> round(qt(.975, 1:40), 3)
 [1] 12.706  4.303  3.182  2.776  2.571  2.447  2.365  2.306
 [9]  2.262  2.228  2.201  2.179  2.160  2.145  2.131  2.120
[17]  2.110  2.101  2.093  2.086  2.080  2.074  2.069  2.064
[25]  2.060  2.056  2.052  2.048  2.045  2.042  2.040  2.037
[33]  2.035  2.032  2.030  2.028  2.026  2.024  2.023  2.021
```

Part II

The experiment described in **Part I** was repeated with a new set of 40 pigs, again arranged in 10 pens of 4 pigs each. In addition to applying the two feed treatments to pens as described in **Part I**, the researchers treated each pig with one of two drugs. Two of the four pigs in each pen were randomly selected for treatment with drug 1, and the other two pigs were treated with drug 2. Each pig was injected intravenously with its assigned drug at the beginning of the three-week period during which feed treatments were applied and weight gains measured. Let y_{ijkl} be the weight gained by lth pig receiving drug k in the jth pen treated with feed type i ($i = 1, 2; j = 1, \ldots, 5; k = 1, 2; l = 1, 2$). Assume the model

$$y_{ijkl} = \mu + \phi_i + \delta_k + \gamma_{ik} + p_{ij} + e_{ijkl}, \tag{2}$$

where μ, ϕ_1, ϕ_2, δ_1, δ_2, γ_{11}, γ_{12}, γ_{21}, and γ_{22} are unknown parameters, $p_{ij} \sim N(0, \sigma_p^2)$ for some unknown variance parameter σ_p^2, $e_{ijkl} \sim N(0, \sigma_e^2)$ for some unknown variance parameter σ_e^2, and all p_{ij} and e_{ijkl} terms are independent. REML estimates of the variance components σ_p^2 and σ_e^2 from the fit of model (2) are

$$\hat{\sigma}_p^2 = 8.5 \quad \text{and} \quad \hat{\sigma}_e^2 = 12.6.$$

Average weight gains are

$$\bar{y}_{1.1.} = 32.7, \;\; \bar{y}_{1.2.} = 31.9, \;\; \bar{y}_{2.1.} = 41.1, \;\; \text{and} \;\; \bar{y}_{2.2.} = 38.7.$$

5. For each of the following linear combinations of model (2) parameters, state whether the linear combination is estimable or not estimable.

a) μ

b) $\mu + \phi_1 + \delta_2 + \gamma_{12}$

c) $\phi_1 - \phi_2$

d) $\phi_1 - \phi_2 + \delta_1 - \delta_2$

e) $\mu + \phi_2$

f) $\mu + \phi_2 + \delta_1/2 + \delta_2/2 + \gamma_{21}/2 + \gamma_{22}/2$

g) $\gamma_{11} - \gamma_{12}$

6. Determine the value of the F statistic you would use to test for a feed-type main effect.

7. State the degrees of freedom associated with the F statistic computed in problem **6**.

8. Determine the value of the F statistic you would use to test for a drug main effect.

9. State the degrees of freedom associated with the F statistic computed in problem **8**.

10. Provide a linear combination of model (2) parameters that is zero if and only if there is no interaction between feed type and drug.

11. Find a 95% confidence interval for the linear combination of parameters provided in problem **10**.

Part III

The experiment described in **Part II** was repeated with a new set of 40 pigs, again arranged in 10 pens of 4 pigs each. The experiment was carried out exactly as described in **Part II** except that weight gain was measured every day for every pig during the three-week treatment period (rather than only at the end of the three-week period). Let y_{ijkld} be the weight gained on day d by lth pig receiving drug k in the jth pen treated with feed type i ($i = 1, 2; j = 1, \ldots, 5; k = 1, 2; l = 1, 2; d = 1, \ldots, 21$). Assume that

$$y_{ijkld} = \mu_{ikd} + p_{ij} + e_{ijkld}, \text{ where} \tag{3}$$

- the μ_{ikd} terms are unknown, real-valued parameters,
- the p_{ij} terms are independent and identically distributed as normal with mean 0 and variance σ_p^2 for some unknown σ_p^2,
- the p_{ij} terms are independent of the e_{ijkld} terms,
- the vectors $\boldsymbol{e}_{ijkl} \equiv (e_{ijkl1}, e_{ijkl2}, \ldots, e_{ijkl20}, e_{ijkl21})'$ are independent and identically distributed as multivariate normal with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}_e$, an unknown 21×21 positive definite matrix.

12. Determine the dimension of the parameter space for model (3).

There are simplified versions of model (3) that can be obtained by making additional assumptions about $\boldsymbol{\Sigma}_e$. Let model (3A) be model (3) with $\boldsymbol{\Sigma}_e = \sigma_e^2 \boldsymbol{A}(\rho)$, where σ_e^2 is an unknown variance component, ρ is an unknown correlation parameter, and $\boldsymbol{A}(\rho)$ is the matrix with $\rho^{|s-t|}$ in row s and column t for all $s, t \in \{1, \ldots, 21\}$. Let model (3B) be model (3) with $\boldsymbol{\Sigma}_e = \nu_e^2 \boldsymbol{B}(\eta)$, where ν_e^2 is an unknown variance component, η is an unknown correlation parameter, and $\boldsymbol{B}(\eta)$ is the matrix with diagonal elements equal to 1 and all other elements equal to η.

13. Assuming model (3A), find an expression for the covariance between the weight gained by a pig on day 1 and the weight gained by that same pig on day 7.

14. Fitting model (3A) to the data yielded the following REML estimates:

$$\hat{\mu}_{111} = 1.99, \hat{\mu}_{117} = 1.69, \hat{\sigma}_p^2 = 0.07, \hat{\sigma}_e^2 = 0.8, \text{ and } \hat{\rho} = 0.9.$$

Provide the value of the t statistic you would use to test $H_0 : \mu_{111} = \mu_{117}$.

15. SAS reports an AIC of 1061.8 when the REML method is used to fit model (3A) to the data. Assuming that model (3A) is more appropriate for the data than model (3B), would you expect the AIC for model (3B) to be less than or greater than 1061.8?

16. Using the REML method in SAS to fit model (3A) to the data resulted in a maximized residual log likelihood of -527.9. Explain how this maximized residual log likelihood value is related to the AIC value of 1061.8 reported by SAS for model (3A).

17. Would you expect the maximized residual log likelihood for model (3) to be less than or greater than -527.9? Explain.

18. Now consider model (3AL), which is the same as model (3A) except that μ_{ikd} is assumed to equal $\beta_{0ik} + \beta_{1ik}d$ for some unknown intercept parameter β_{0ik} and some unknown slope parameter β_{1ik} for all i, k, and d. Using the ML method in SAS to fit model (3AL) to the data resulted in a maximized log likelihood of -490. Using the ML method in SAS to fit model (3A) to the data resulted in a maximized log likelihood of -435. Based on these maximized log likelihood values, does it seem like model (3AL) provides an adequate fit to the data relative to model (3A)? Explain your reasoning using simple calculations to support your answer.

Part IV

A slackline is a strip of webbing, typically no more than two inches wide and at least 40 feet long, that can be anchored at each end and suspended above the ground. A person can walk across the line in a manner similar to a tightrope walker. A sample of 500 people with slacklining experience each attempt to walk from one end of a 50 foot slackline to the other. Let $x_i \in (0, 50]$ be the distance the ith person was able to cover along the slackline before falling. Let $y_i \in (0, 1]$ be $x_i/50$, i.e., the proportion of the length of the slackline successfully traversed by the ith person. Note that $y_i = 1$ implies that the ith person was able to cross the slackline successfully from end to end.

Suppose that $y_1, \ldots, y_{500}$ are independent and identically distributed. Suppose $y_i = 1$ with probability π, and $y_i < 1$ with probability $1 - \pi$. Suppose the conditional distribution of y_i given $y_i < 1$ is Beta$(\theta, 5)$ for some parameter $\theta > 0$, which implies $E(y_i|y_i < 1) = \frac{\theta}{\theta+5}$.

19. Suppose 150 of the 500 people successfully traversed the slackline from end to end. Provide a confidence interval for π with coverage probability approximately equal to 0.95.

20. Without loss of generality, use $i = 1, \ldots, 350$ to index the people who did not successfully traverse the slackline from end to end. Suppose the MLE of θ from $y_1, \ldots, y_{350}$ is $\hat{\theta} = 4.0$. Now imagine randomly selecting a person from the same population as the 500 people in our sample. Let λ be the probability that the randomly selected person makes it at least 75% of the way across the slackline. Find an expression for a confidence interval for λ with coverage probability approximately equal to 0.95. Your answer is allowed to depend on $G(y; \theta)$, the CDF of the Beta$(\theta, 5)$ distribution, as well as derivatives of $G(y; \theta)$ with respect to y, θ, or both. Use $G^{(u,v)}(y; \theta)$ to denote the uth derivative with respect to y and the vth derivative with respect to θ of $G^{(u,v)}(y; \theta)$. For example, $G^{(1,0)}(0.75; 4)$ is the first derivative $G(y; \theta)$ with respect to y evaluated at $y = 0.75$ and at $\theta = 4$ (the MLE of θ).

1 Background

There are many factors that influence the vote in U.S. presidential elections, and these factors continue to be the focus of various statistical (and some not-so-statistical) analyses. The two major political parties in the United States are called the Democratic and Republican parties. Although there may be a number of independent candidates or candidates from other parties (such as the Green Party) we will focus here only on the two major parties. One belief that seems to be widely accepted is that higher voter turnout favors the candidate that is a Democrat over the candidate that is a Republican. The development of a statistical model to examine this belief is the topic to be considered in this question.

For a variety of reasons that are crucial to the actual analysis of this problem but not so important for this question on the written preliminary examination, we will consider the development of models to relate variables defined as the change from 2012 to 2016 in the proportion of votes cast for the Democrat and the change from 2012 to 2016 in the proportion of registered voters that cast votes. Our sampling units will be states of the United States, including also the District of Columbia (Washington D.C.) for a total of 51 observations. Because of four missing values in the proportion of registered voters who voted, that total is reduced to 47 observations. So define the random variables $Y_i;\ i = 1, \ldots, n$ as associated with the change in the proportion of votes cast for the Democrat between 2016 and 2012 in state i, and the random variables $X_i;\ i = 1, \ldots, n$ as the change in the proportion of registered voters who actually voted between 2016 and 2012 in state i. Both of these changes are taken as the value for 2016 minus the value for 2012. A scatterplot of the data in question is presented in Figure 1. Notice from this plot that both axes are fairly compressed relative to the theoretically possible values for a difference in proportions of $(-1, 1)$. Also, while there may be some indication of a positive relation in Figure 1, what would be considered the "signal to noise ratio" is low.

2 Developing a Regression Model

Skipping over questions related to the development of correlation versus regression models, suppose we are willing to condition on the values of X_i observed and treat these values as fixed covariates in a regression model. Thus, we desire to develop a regression model to relate the distributions of the Y_i to the covariates x_i. One way to begin the model

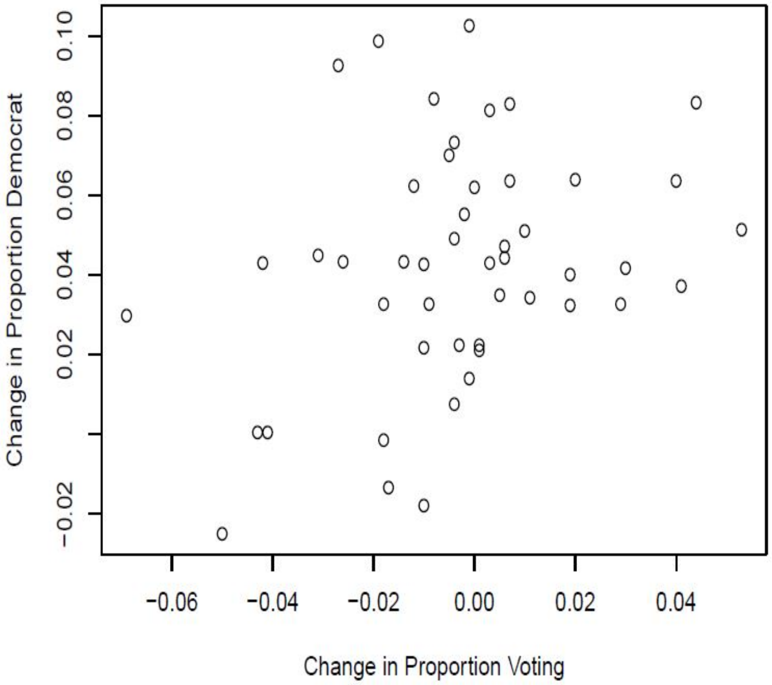


Figure 1: Scatterplot of proportional difference in votes for the Democrat versus proportion voting .

development process is to consider what type of a random component we might choose for the responses at a given covariate value. The visual impression based on Figure 1 is that there is not a great deal of information in the data to guide this choice; there are only 47 values in total, and the number of responses are not evenly spread over the range of the covariate values, making examination of empirical distributions for binned values of limited use. We do know that the variables used to construct the Y_i are proportions in different years, so the range of possible values is $(-1, 1)$.

Dealing with proportions suggests consideration of beta distributions. There are many number of generalizations of what we consider a traditional beta distribution (as seen in Stat 542), some of which do have support on part of the negative line as well as part of the positive line. But none of these existing options seem to be easily developed into a suitable model for our objective.

Consider an entirely ad hoc possibility consisting of a linear combination of two beta distributions, one scaled to the negative unit interval and the other a traditional beta. If X has a traditional beta distribution with parameters $\alpha > 0$ and $\beta > 0$, let $Z = -X$. The density of Z is,

$$f_z(z|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}|z|^{\alpha-1}(1-|z|)^{\beta-1}; \quad -1 < z < 0.$$

If we would consider negative values of our responses Y_i to have this distribution and positive values a regular beta distribution, then the distribution of a randomly selected response would be,

$$f(y|\alpha_n,\beta_n,\alpha_p,\beta_p,\gamma) = \begin{cases} (1-\gamma)K_n|y|^{\alpha_n-1}(1-|y|)^{\beta_n-1} & \text{for } -1 < y < 0 \\ \gamma K_p y^{\alpha_p-1}(1-y)^{\beta_p-1} & \text{for } 0 < y < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where $K_n = \Gamma(\alpha_n)/\Gamma(\alpha_n+\beta_n)$ and $K_p = \Gamma(\alpha_p)/\Gamma(\alpha_p+\beta_p)$. An implication of (1) is that $\gamma = Pr(Y > 0)$.

A simplification that seems reasonable for our application is to take $\alpha_n = \alpha_p$ and $\beta_n = \beta_p$. For a given state in the U.S., it is not unreasonable to assume that the distribution of our response for an increase in voter turnout should be similar to the distribution for a decrease in voter turnout, but with opposite sign on the response itself. This restriction of parameters gives that the portion of the density (1) on the negative line will be a mirror image of the portion on the positive line. For lack of a better name, then, we might call (1) a *two-faced beta distribution.*

There remains the potential difficulty caused by the fairly small range of responses exhibited by our data. This can be dealt with using a simple scale transformation. Again, it seems within the scope of reason that the negative and positive ranges should be the same. If X has a two-faced beta distribution with density (1), let $Y = \lambda X$ for some $0 < \lambda < 1$. This results in the following two-faced beta distribution, for $\alpha > 0$, $\beta > 0$, $0 < \gamma < 1$, and $0 < \lambda < 1$,

$$f_Y(y|\alpha,\beta,\gamma,\lambda) = \begin{cases} K(1-\gamma)\left(\frac{1}{\lambda}\right)^{\alpha+\beta-1}|y|^{\alpha-1}(\lambda-|y|)^{\beta-1} & \text{for } -\lambda < y < 0, \\ K\gamma\left(\frac{1}{\lambda}\right)^{\alpha+\beta-1}y^{\alpha-1}(\lambda-y)^{\beta-1} & \text{for } 0 < y < \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

where $\gamma = Pr(Y > 0)$ and,

$$K = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

To get some feeling for this distribution, several densities having different parameter values are graphed in Figure 2. Note that the negative and positive faces of these distributions are not exactly symmetric because $\gamma = 0.60$ in each case, and the range of values is -0.15 to 0.15 becasue $\lambda = 0.15$ in each case.

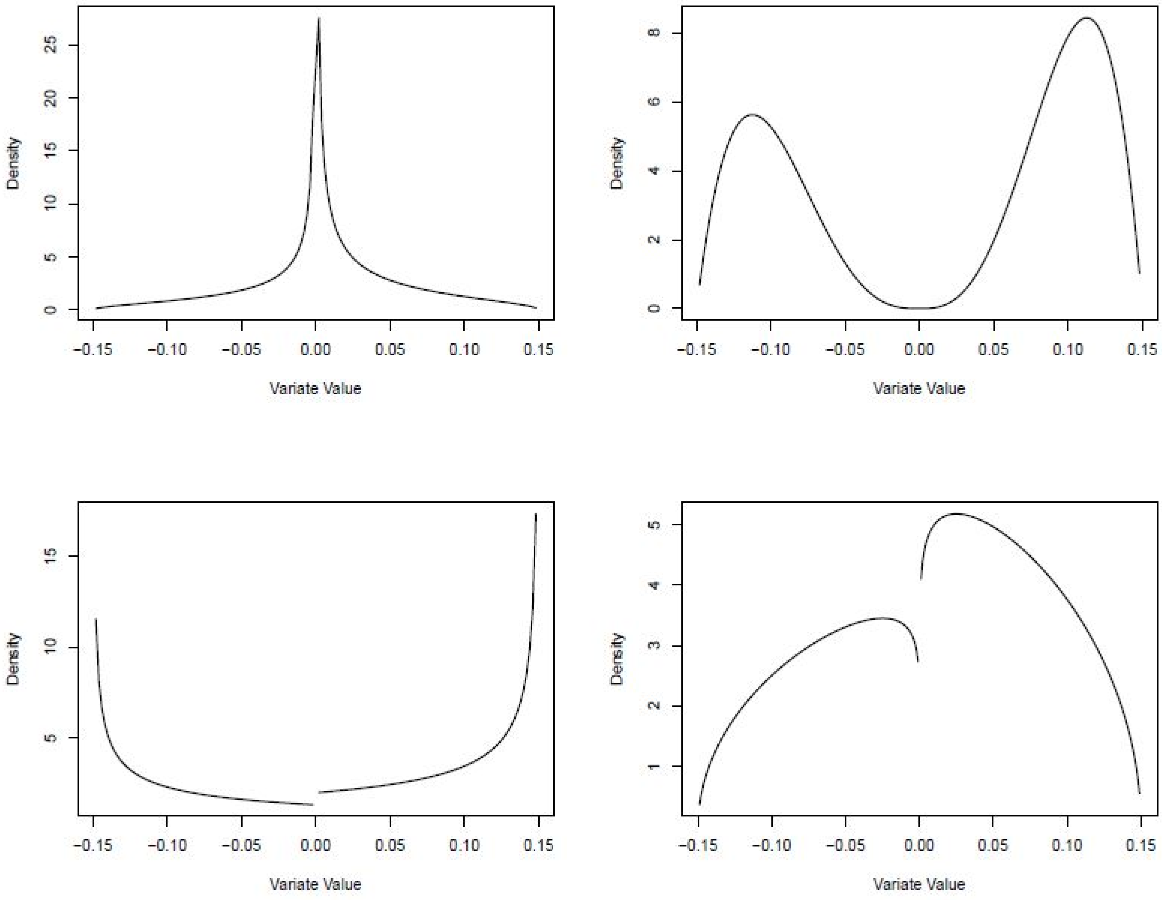


Figure 2: Several two-faced beta density functions. In these densities, $\gamma = 0.6$ and $\lambda = 0.15$ in each case. For the density in the upper left, $\alpha = 0.35$ and $\beta = 1.5$. For the density in the upper right, $\alpha = 4$ and $\beta = 2$. For the density in the lower left, $\alpha = 1$ and $\beta = 0.5$, and for the density in the lower right, $\alpha = 1.1$ and $\beta = 1.5$.

Question 1. If Y is a random variable with probability density function given by (2), find $E(Y)$ and $var(Y)$ in terms of μ, ϕ, γ and λ. Let $\mu = \alpha/(\alpha+\beta)$ and $\phi = 1/(\alpha+\beta+1)$ be the usual expressions used in a mean value parametrization of a beta distribution. Can the variance of Y be written as proportional to some function of $E(Y)$?

Hint 1: It is probably easiest to find the expected value and variance for a random variable Z that has density (1) and then make the transformation $Y = \lambda Z$.

Question 2. Demonstrate that the effect of a covariate on $E(Y)$ should be incorporated into a model for γ in (2), not $\mu = \alpha/(\alpha+\beta)$. That is, argue that if (2) is used as a distribution for Y_i, we should allow γ_i to vary across observations, but keep μ and ϕ (and thus also α and β) constant across observations.

Hint: Consider that in our regression, we want the expected values of the Y_i to be both negative and positive.

The parameter space for γ in (2) is $[0, 1]$. If we allow this parameter to vary across observations, we might then choose a link function $g(\cdot)$ that has this range. A natural possibility is a logit link for which, technically, $\gamma \in (0, 1)$, and

$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right) = \psi_0 + \psi_1 x_i, \tag{3}$$

where ψ_0 and ψ_1 are unknown regression parameters and x_i is the change from 2012 to 2016 in the proportion of registered votes actually casting votes in state i. Our two-faced beta regression model is then formulated for independent random variables $Y_1, \ldots, Y_n$ having density functions $f_Y(\alpha, \beta, \gamma_i, \lambda)$ as in (2) with γ_i further specified as in (3).

3 Estimation and Inference

3.1 A One-Sample Problem

To approach the tasks of estimation and inference for our model we might first consider a simpler problem of estimating the parameters based on a random sample. If $Y_1, \ldots, Y_n$ are independent and identically distributed according to a distribution with density (2), consider estimation of the parameters α, β, γ and λ. Initially, we will consider estimation via maximum likelihood.

Question 3. Write the log likelihood that results from a random sample of the distribution with density (2) and parameters $\alpha > 0$, $\beta > 0$, $0 < \gamma < 1$, and $0 < \lambda \leq 1$.

Question 4. Before worrying about the other parameters, notice that the maximum likelihood estimator of γ is independent of all other parameters and may be found in closed form. Find this estimator.

Question5. Because of the result of question 4, we have reduced the problem to finding estimators for α, β, and λ. Demonstrate (show) that estimation of these parameters from the actual sample will be the same problem as estimation of the parameters of a scaled beta distribution based on the absolute values of our actual sample. That is, show that maximum likelihood estimation of α, β, and λ based on a random sample from (2) is identical to estimation of those same parameters based on the absolute values considered to be a random sample from that distribution with $\gamma = 1$.

If $\gamma = 1$ in (2) the density becomes

$$f_Y(y|\alpha,\beta,\lambda) = \begin{cases} K\left(\frac{1}{\lambda}\right)^{\alpha+\beta-1} y^{\alpha-1}(\lambda-y)^{\beta-1} & \text{for } 0 < y < \lambda \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We will now consider estimation of α, β, and λ based on a random sample from a scaled beta distribution given by (4). Such estimation is potentially complicated by the relation between the support of (4) and the parameter space for λ. In particular, without any data the parameter space of λ is $(0, 1]$. With data in hand, however, we know that λ can be no smaller than the the maximum observed value. If we would take $\alpha = \beta = 1$ the problem would further reduce to estimating λ from a Unif$(0, \lambda)$ distribution and we could take $\hat{\lambda}$ to be the largest order statistic, namely,

$$\hat{\lambda} = \max\{Y_1, \ldots, Y_n\} = Y_{[n]}.$$

It turns out that for many sets of data the behavior of a log likelihood based on a random sample from (4) is similar to that of the uniform problem, in that the profile log likelihood in λ is strictly increasing as λ decreases to the maximum value in the sample. There is a bit of a distinction with the uniform problem, however, when $\beta \neq 1$ that could cause some additional difficulty here if we attempt to take the estimator of λ to be $y_{[n]}$.

Question 6. Identify the difficulty just described. Why is it a problem to take $\hat{\lambda} = y_{[n]}$ when $\beta \neq 1$ but not when $\beta = 1$?

Hint: Consider what happens to the likelihood for the uniform problem if we take $\lambda = y_{[n]}$ and what happens to the likelihood (or log likelihood) for our current problem if we take $\lambda = y_{[n]}$.

Even if we could define the maximum likelihood estimator of λ to be the largest order statistic and somehow circumvent the difficulty you identified in question 6, we know that this estimator would have negative bias and must be smaller than the actual value of λ. It turns out that, for a fixed α and β, the expected value of the largest order statistic in a random sample from (4) is monotone in λ. Now, the maximum likelihood estimates of α and β are (at least numerically) available for a fixed value of λ as $\alpha(\lambda)$ and $\beta(\lambda)$ and we suppose that these will not change greatly for small changes of λ. Suppose that the expected value of the largest order statistic can be expressed as a function $h(\lambda, \alpha(\lambda), \beta(\lambda))$ such that $h(\lambda, \alpha(\lambda), \beta(\lambda))$ is a monotone function of λ in some neighborhood of the value of λ that makes the expected value of $Y_{[n]}$ equal to $y_{[n]}$. If this is the case, an estimation procedure is suggested as follows:

1. For a given set of data presumed to be a random sample of size n from (4), determine the maximum value in the sample, $y_{[n]} = \max\{y_i : i = 1, n\}$.

2. Form a transect of values of λ_m increasing from $y_{[n]} + \delta$ to $y_{[n]} + M\delta$ for some small δ and integer M, that is, $\lambda_m = y_{[n]} + m\delta;\ m = 1, \ldots, M$.

3. Compute estimates $\alpha(\lambda_m)$ and $\beta(\lambda_m)$ by maximizing the log likelihood for fixed λ_m, for $m = 1, \ldots, M$.

4. For each set of parameters λ_m, $\alpha(\lambda_m)$, $\beta(\lambda_m)$, compute a numerical approximation to the expected value of the largest order statistic $E(Y_{[n]}|\alpha(\lambda_m), \beta(\lambda_m), \lambda_m)$, using the actual sample size, n. Call these approximations $E_m;\ m = 1, \ldots, M$.

5. Choose a value λ_L for which E_L is the largest value of E_m less than $y_{[n]}$, and a value λ_U for which E_U is the smallest value of E_m greater than $y_{[n]}$. If these values are not available, return to step 2 and adjust the transect.

6. Using λ_L and λ_U as a bracket, determine the value of λ that produces an expected largest order statistic equal to $y_{[n]}$, using a bisection algorithm, say. Take this value to be the estimate of λ, $\hat{\lambda}$, say. Take the estimates of α and β to be $\hat{\alpha} = \alpha(\hat{\lambda})$ and $\hat{\beta} = \beta(\hat{\lambda})$.

This algorithm produces what we might consider a moment-based estimator of λ, since it equates the expected value of $Y_{[n]}$ with the observed maximum value in the sample, $y_{[n]}$, and conditional maximum likelihood estimates of α and β given that $\lambda = \hat{\lambda}$. For purposes of reference, we will call this the ABL algorithm (for α, β, λ).

3.2 The Regression Problem

Turn now to the full problem under consideration in which we have independent random variables $Y_1, \ldots, Y_n$ from a distribution with densities, for $i = 1, \ldots, n$, $\alpha > 0$, $\beta > 0$, $0 < \lambda < 1$, and $0 < \gamma_i < 1$,

$$f_Y(y|\alpha, \beta, \gamma_i, \lambda) = \begin{cases} K(1-\gamma_i)\left(\frac{1}{\lambda}\right)^{\alpha+\beta-1} |y|^{\alpha-1} (\lambda - |y|)^{\beta-1} & \text{for } -\lambda < y < 0, \\ K\gamma_i \left(\frac{1}{\lambda}\right)^{\alpha+\beta-1} y^{\alpha-1} (\lambda - y)^{\beta-1} & \text{for } 0 < y < \lambda, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

where

$$K = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

and,

$$\log\left(\frac{\gamma_i}{1-\gamma_i}\right) = \psi_0 + \psi_1 x_i.$$

At this juncture, we wish to determine whether estimation will follow a pattern similar to what we have determined for the one-sample problem.

Question 7. Consider, first, estimation of the regression parameters ψ_0 and ψ_1. Draw on what you developed in question 4 to describe how we can estimate these parameters independently of α, β, and λ.

In the regression model, the outcome of question 5 continues to hold, and the parameters α, β and λ can be estimated by considering the absolute values of responses as arising from a random sample from (4), using the ABL algorithm outlined in the previous section. An overall likelihood-based estimation procedure is then to estimate ψ_0 and ψ_1 following your answer to Question 7, and then estimate the other parameters using the ABL algorithm, after replacing responses with their absolute values. To illustrate this procedure, consider the simulated data presented in Figure 3, which contains 51 observations and has an observed correlation between the covariate and response of 0.52, a little higher than in the actual data. A summary of estimation results, along with parameter values used to simulate the data, are given in Table 1. In these data, the maximum absolute response was 0.1165. In Step 5 of the ABL algorithm, it was determined that $\lambda^{(0)} = 0.15$ produced an expected largest order statistic of 0.1143, less than the observed maximum, while a value of $\lambda^{(1)} = 0.17$ produced an expected largest order statistic of 0.1177, greater than the observed maximum. Bracketing the solution with these two values of λ, Step 6 of the ABL algorithm, using bisection, returned $\hat{\lambda}$ in Table 1 as the estimated value of λ. The profile log likelihood at this value produced the estimates $\hat{\alpha}$ and $\hat{\beta}$ in Table 1.

Parameter	True Value	Estimate
ψ_0	0.0	-0.067
ψ_1	20	25.686
α	0.75	0.954
β	1.50	4.190
λ	0.12	0.1623

Table 1: Estimates and generating values of parameters for data in Figure 3.

Quantities of interest in the problem under consideration are the γ_i as a function of

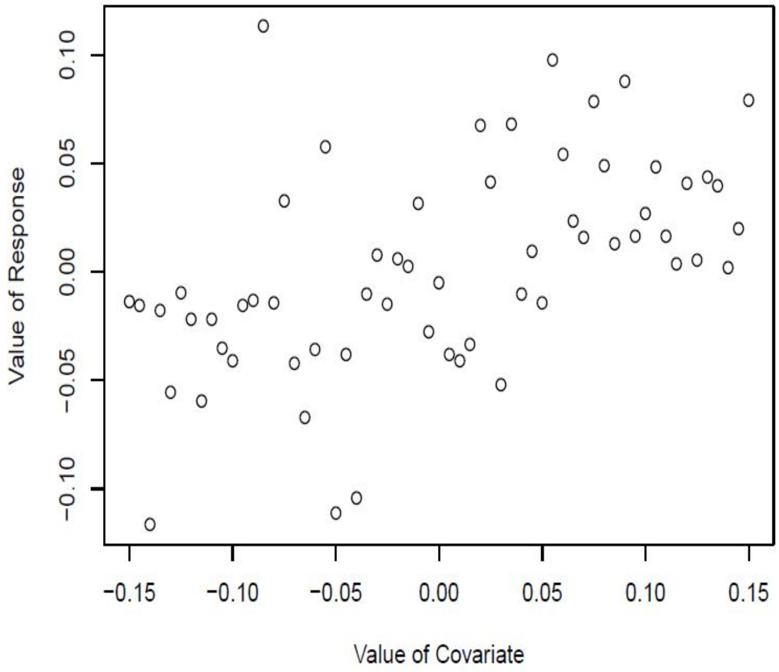


Figure 3: Scatterplot of data simulated from the two-faced beta regression model (5).

the covariate values x_i, the underlying two-faced beta distribution with parameters α, β and λ and, to a lesser extent, the expected responses, $E(Y_i)$. Figure 4 contains a plot of the estimated and true values of

$$\gamma_j = \frac{\exp(\eta_j)}{1 + \exp(\eta_j)},$$

where $\eta_j = \psi_0 + \psi_1 x_j$ for x_j. Graphs of estimated response densities for four particular values of γ_i are given in Figure 5, while the true densities used to generate the data are shown in Figure 6. Finally, the scatterplot of Figure 3 is reproduced with the true and estimated expectation functions overlain in Figure 7.

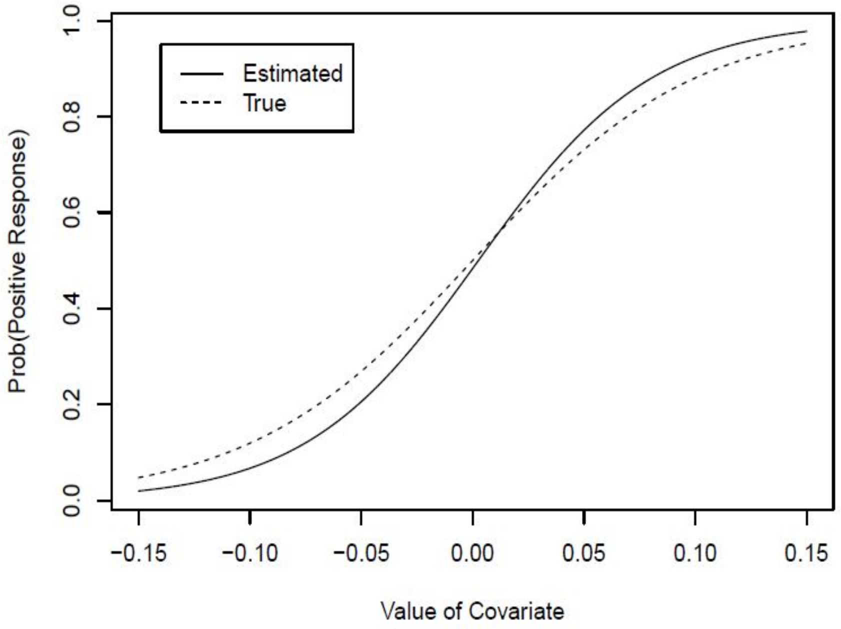


Figure 4: Estimated (solid curve) and True (dashed curve) values of γ_i as a function of covariate values x_i for the simulated data set.

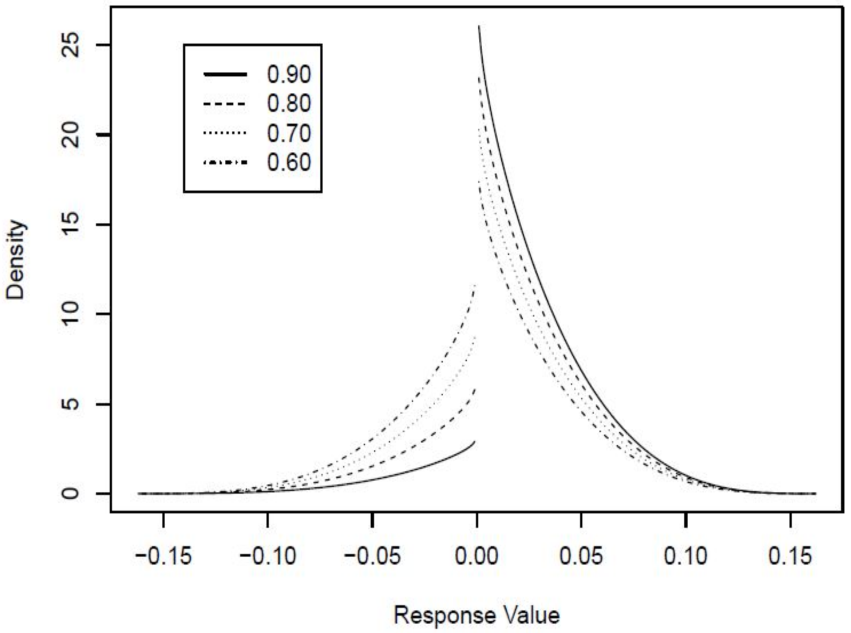


Figure 5: Estimated response densities at various values of γ_i for the simulated data.

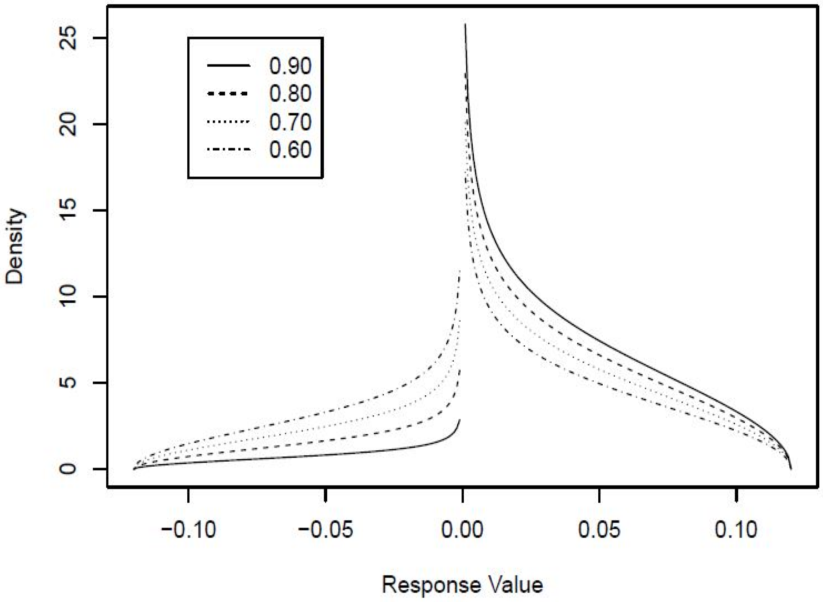


Figure 6: True response densities at various values of γ_i for the simulated data.

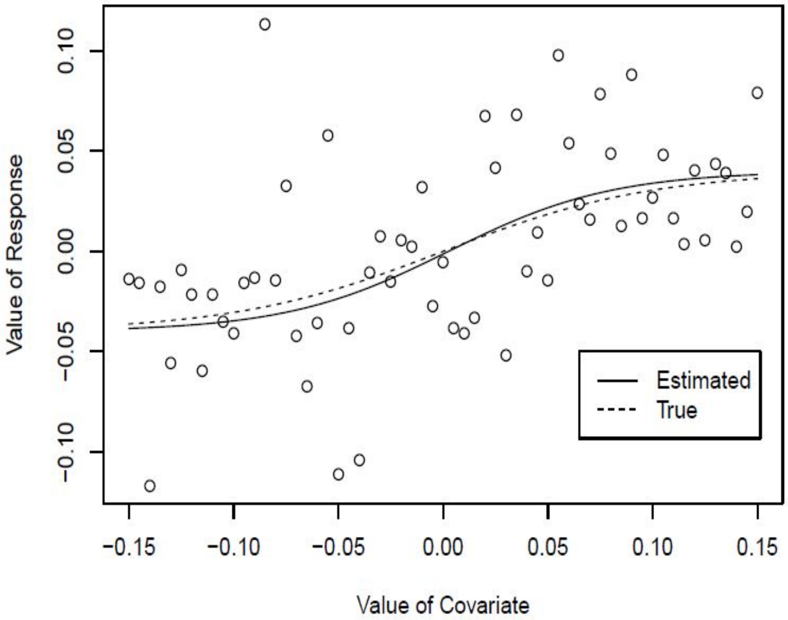


Figure 7: Simulated data along with estimated (solid curve) and true (dashed curve) expectation functions.

Question 8. It is not clear how we should approach the task of producing inferential statements about our fitted model. Methods that come to mind are (i) computation of intervals based on asymptotic normality of maximum likelihood estimators (Wald Theory), (ii) profile likelihood and associated intervals, or (iii) parametric bootstrap. For each of three sets of parameters, $\{\psi_0, \psi_1\}$, $\{\alpha, \beta\}$, and $\{\lambda\}$, comment on why each of these possibilities may or may not be applicable for computing interval estimates.

Note: This is a rather major question and requires a more extensive answer than many of the specific questions asked. A complete solution will discuss all three inference methods for each parameter set and thereby have 9 total parts.

3.3 Bayesian Considerations

Consider now the same model of expression (5) but suppose we desire to conduct an analysis using Bayesian methods.

Question 9. Suggest a change to the notation of the model (but not the model itself) that would lead to obvious and simple prior distributions for the parameters of the beta distribution, excluding λ.

Hint: Do not over-think this.

Question 10. In the likelihood estimation problem, the regression parameters and the other parameters separated into independent pieces. Will this same phenomenon benefit us in simulation from the joint posterior distribution? Determine the answer by writing out the full conditional posterior distributions of ψ_0, ψ_1, α, β, and λ for use in a Gibbs Sampling algorithm. In doing this, assume that the joint prior has been specified in a product form as

$$\pi(\psi_0, \psi_1, \alpha, \beta, \lambda) = \pi(\psi_0)\,\pi(\psi_1)\,\pi(\alpha)\,\pi(\beta)\,\pi(\lambda).$$

Hint: Note that the joint data density can be written as

$$\begin{aligned} f_Y(\boldsymbol{y}|\psi_0, \psi_1, \alpha, \beta, \lambda) &= \prod_{y_i \in C_n} (1-\gamma_i) K \frac{1}{\lambda^{\alpha+\beta-1}} |y_i|^{\alpha-1} (\lambda - |y_i|)^{\beta-1} \\ &\times \prod_{y_i \in C_p} \gamma_i K \frac{1}{\lambda^{\alpha+\beta-1}} y_i^{\alpha-1} (\lambda - y_i)^{\beta-1} \end{aligned}$$

where C_n denotes the set of observations that are negative, C_p denotes the set of observations that are positive, and $K = \alpha/(\alpha+\beta)$ as previously.

NOTE: This model has been fitted to the actual data, and some model diagnostics run to demonstrate that it appears appropriate for describing the data. The results are kind of interesting, and if you have interest feel free to ask after the exam is over.

Part I

1. This experiment is from an RCBD (matched pair) design. There are three different non-parametric methods covered in 500 for such designs.

Block	Schedule 1	Schedule 2	Difference	Rank	Sign
1	34.98	37.18	2.2	2	+
2	41.22	45.85	4.63	4	+
3	36.94	40.23	3.29	3	+
4	39.97	39.20	-0.77	1	-

(1) the Wilcoxon signed rank test

For the observed data, W = 2+3+4 = 9, more extreme cases in the same direction are W = 1+ 2+3+4 = 10. The exact p-value is 2x2/16 = 0.25.

(2) the sign test

The exact p-value = 2 x $\left[\binom{4}{1}0.5^4 + \binom{4}{0}0.5^4\right]$ = 0.625.

For both non-parametric tests, we fail to detect any difference between nitrate content due to difference schedules.

Part II

2.

Source	DF
Block	3
Schedule	1
FertilizerAmount	2
Schedule*FertilizerAmount Interaction	2
Error	15
Corrected Total	23

3. The baseline constraints in SAS set the following parameters to be 0:
α_2, γ_3,, $(\alpha\gamma)_{21}$, $(\alpha\gamma)_{22}$, $(\alpha\gamma)_{23}$, $(\alpha\gamma)_{13}$

4. The parameter vector is: $[\mu, \alpha_1, \gamma_1, \gamma_2, (\alpha\gamma)_{11}, (\alpha\gamma)_{12}]^T$

The first 6 rows of the design matrix X is given below:

$$\begin{vmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{vmatrix}$$

5. α_1 represents the difference between schedule 1 and schedule 2 when fertilizer amount is set at level high, i.e., $\alpha_1 = \mu_{13} - \mu_{23}$.
$(\alpha\gamma)_{12}$ represents the interaction effect between schedule and fertilizer amount:

$(\alpha\gamma)_{12} = \mu_{12} - \mu - \alpha_1 - \gamma_2 = \mu_{12} - (\mu + \alpha_1) - (\mu + \gamma_2) + \mu = \mu_{12} - \mu_{13} - \mu_{22} + \mu_{23}$
$= (\mu_{12} - \mu_{13}) - (\mu_{22} - \mu_{23})$

6. The null hypothesis is $(\alpha\gamma)_{11} = (\alpha\gamma)_{12} = 0$.

7. The F-statistic = 1.00/7.20 = 0.14, and the degrees of freedom for this F-statistic is (2, 15).

8. The difference in the mean response between the two schedules when fertilizer amount is low is a simple effect.

9. Depending on whether we test the simple effect directly or test for the main effect (because of no interaction between the two treatment factors), there are two possible tests based on model (1).
(1) Test the simple effect directly:
The point estimate of the simple effect is 2.3375.
The standard error for the estimate is $\sqrt{MS_{Error}\left({}^{2}/_{4}\right)} =$ 1.897. Then

$$t = {}^{2.3375}/_{1.897} = 1.232$$

p-value = 2*P(t_{15} > 1.232) = 0.237

Based on the p-value, we fail to reject the null hypothesis and conclude that there is no significant difference in the mean response between the two schedules when fertilizer amount is low.

(2) Test the main effect:
The point estimate of the simple effect is 2.93.
The standard error for the estimate is $\sqrt{MS_{Error}\left({}^{2}/_{12}\right)} =$ 1.095. Then

$$t = {}^{2.93}/_{1.095} = 2.676$$

p-value = 2*P(t_{15} > 2.676) = 0.017

Based on the p-value, we reject the null hypothesis and conclude that there is significant difference in the mean response between the two schedules. Note that the test for the main effect is more powerful than the test for the simple effect due to more precise estimate of the effect.

10. The normality assumption is reasonable based on the normal probability plot and histogram. The equal variance assumption is also appropriate based on the residual plots.

11. Since the model assumptions are okay for model (1), the analysis from problem 9 is preferred as it is more powerful.

12. The null hypothesis is $\gamma_1 + 0.5(\alpha\gamma)_{11} = \gamma_2 + 0.5(\alpha\gamma)_{12} = 0$.
The F-statistic = 73.90/7.20 = 10.26.
The total SS can be calculated by SSE and R2 given on page 5 (SSTotal = SSE/(1-R2)).
Then the SS for fertilizer can be calculated and hence MS for fertilizer can be calculated as well.
The degrees of freedom for the corresponding F distribution is (2, 15).

13. Based on model (1), $Var(Y_{ijk}) = \sigma_\beta^2 + \sigma^2$
Based on the following ANOVA table, the estimates for σ^2 and σ_β^2 are, respectively, 7.20 and (65.67-7.20)/6 = 9.745
Hence, the estimated value of $Var(Y_{ijk})$ is 7.20+9.745 = 16.945

Type 3 Analysis of Variance

Source	DF	Sum of Squares	Mean Square	Expected Mean Square
block			65.67	Var(Residual) + 6 Var(block)
Residual			7.20	Var(Residual)

14. The estimated correlation between Y_{ijk} and Y_{ijl}, where $k \neq l$, is 9.745/16.945 = 0.575.

15. The correlation between Y_{ijk} and Y_{mjk} is 0.

16. The complete ANOVA table is:

Source	**DF**
Block	3
Schedule	1
FertilizerAmount	2
Schedule*FertilizerAmount Interaction	2
Block* Schedule	3
Block* FertilizerAmount	6
Block* Schedule*FertilizerAmount Interaction	6
Error	24
Corrected Total	47

Part III

17. The estimated value of parameter β_1 can be interpreted as:
For plants with the same height, leaf area, soil nitrate concentration, and soil potassium concentration in the middle of growing period, the mean fresh weight for plants under well-watered condition is 9.2029 grams heavier than the mean fresh weight for plants under water-stress condition.

18. Based on the diagnostic plots for model (2), there are no obvious deviation from normality and equal variance assumptions.

19. H$_0$: $\beta_4 = \beta_5 = 0$ vs H$_a$: At least one of β_4 and β_5 is not equal to 0.

$F = \frac{(SSE_{model3} - SSE_{model2})/2}{MSE_{model2}} = \frac{(2.1-0.48)/2}{0.12}$ =6.75

The degrees of freedom for the corresponding F distribution is (2, 4).

20. We can use several criteria to compare the three models. Students may base their conclusion on any criterion as long as they justify accordingly. In fact, based on all criteria, model (2) is the best.

model	Adjusted R^2	Cp	AIC	BIC
2	0.9945	6	-18.41	-16.60
3	0.9838	15.628	-7.58	-6.37
4	0.9904	8.46	-12.80	-11.59

Adjusted $\mathrm{R}^2 = 1 - \frac{MS_{error}}{SS_{total}/_{(n-1)}}$

$\mathrm{Cp} = \frac{SS_{error}}{\widehat{\sigma^2}} - (n - 2(k+1))$

$\mathrm{AIC} = \mathrm{n} \log\left(\frac{\mathrm{SSerror}}{\mathrm{n}}\right) + 2(\mathrm{k} + 1)$

$\mathrm{BIC} = \mathrm{n} \log\left(\frac{\mathrm{SSerror}}{\mathrm{n}}\right) + (\mathrm{k} + 1)\log(n)$

21. To accommodate that the relationship between soil nutrients and fresh weight is dependent on the water conditions, we include the interaction between water condition and soil nutrients:

$$Y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{1k} x_{4k} + \beta_7 x_{1k} x_{5k} + \varepsilon_k$$

1. $\boldsymbol{X} = [\mathbf{1}_{40\times 1}, \boldsymbol{I}_{2\times 2} \otimes \mathbf{1}_{20\times 1}]$, $\boldsymbol{Z} = \boldsymbol{I}_{10\times 10} \otimes \mathbf{1}_{4\times 1}$

2. The ANOVA table with expected mean squares is straightforward in this case because of the balanced design. We have

Source	DF	Sum of Squares	Expected Mean Squares
$feed$	$2-1$	$5\cdot 4\cdot \sum_{i=1}^{2}(\overline{y}_{i\cdot\cdot} - \overline{y}_{\cdot\cdot\cdot})^2$	$\sigma_e^2 + 4\sigma_p^2 + \frac{5\cdot 4}{2-1}\ \sum_{i=1}^{2}(\phi_i - \overline{\phi}_\cdot)^2$
$pen(feed)$	$2\cdot(5-1)$	$4\cdot \sum_{i=1}^{2}\sum_{j=1}^{5}(\overline{y}_{ij\cdot} - \overline{y}_{i\cdot\cdot})^2$	$\sigma_e^2 + 4\sigma_p^2$
$pig(pen, feed)$	$2\cdot 5\cdot(4-1)$	$\sum_{i=1}^{2}\sum_{j=1}^{5}\sum_{k=1}^{4}(y_{ijk} - \overline{y}_{ij\cdot})^2$	σ_e^2
$c.total$	$2\cdot 5\cdot 4 - 1$	$\sum_{i=1}^{2}\sum_{j=1}^{5}\sum_{k=1}^{4}(y_{ijk} - \overline{y}_{\cdot\cdot\cdot})^2$	

Thus, $\hat{\sigma}_p^2 = (52.252 - 10.011)/4 \approx 10.56$ is an unbiased method-of-moments estimator and the REML estimator of σ_p^2.

3. Because pens are the experimental units to which the levels of the factor feed type were assigned, and because we have a balanced design, the appropriate error term for testing for feed type effects is pens (nested within feed types). Thus, the relevant statistic is $F = 119.025/52.252 \approx 2.28$.

4. The BLUE of $\phi_1 - \phi_2$ is $\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}$, and

$$\mathrm{Var}(\bar{y}_{1\cdot\cdot} - \bar{y}_{2\cdot\cdot}) = \mathrm{Var}(\bar{p}_{1\cdot} - \bar{p}_{2\cdot} + \bar{e}_{1\cdot\cdot} - \bar{e}_{2\cdot\cdot}) = 2\sigma_p^2/5 + 2\sigma_e^2/20 = \frac{1}{10}EMS_{\text{pen(feed)}}.$$

Thus, the appropriate confidence interval is

$$\begin{array}{rcl} 31.69 - 35.14 & \pm & t_{.975,8}\sqrt{0.1MS_{\text{pen(feed)}}} \\ -3.45 & \pm & 2.306\sqrt{5.2252} \\ -3.45 & \pm & 5.27 \Longleftrightarrow (-1.82, 8.72). \end{array}$$

5. A linear combination is estimable if and only if it is a linear combination of the expected values of the elements in the response vector. In this case, each element of the response vector has one of four expected values:

$$\mu + \phi_1 + \delta_1 + \gamma_{11}, \mu + \phi_1 + \delta_2 + \gamma_{12}, \mu + \phi_2 + \delta_1 + \gamma_{21}, \mu + \phi_2 + \delta_2 + \gamma_{22}.$$

It is straightforward to show that only **b** and **f** can be written as linear combinations of the four cell means. Thus, only the linear combinations in **b** and **f** are estimable.

6. The feed-type main effect is $\phi_1 - \phi_2 + \bar{\gamma}_{1.} - \bar{\gamma}_{2.}$, whose BLUE is $\bar{y}_{1...} - \bar{y}_{2...}$. We have

$$\text{Var}(\bar{y}_{1...} - \bar{y}_{2...}) = \text{Var}(\bar{p}_{1.} - \bar{p}_{2.} + \bar{e}_{1...} - \bar{e}_{2...}) = 2\sigma_p^2/5 + 2\sigma_e^2/20.$$

Thus, a t statistic for testing for a feed-type main effect is

$$\frac{\bar{y}_{1...} - \bar{y}_{2...}}{\sqrt{2\hat{\sigma}_p^2/5 + 2\hat{\sigma}_e^2/20}} = \frac{32.3 - 39.9}{\sqrt{3.4 + 1.26}} \approx -3.52,$$

which implies that the F statistic is approximately $3.52^2 \approx 12.39$.

7. This is a split-plot experiment with multiple observations per split-plot experimental unit. The relevant sources of variation and their degrees of freedom are as follows.

Source	Degrees of Freedom
Feed	$2 - 1 = 1$
Pen(Feed)	$2(5 - 1) = 8$
Drug	$2 - 1 = 1$
Feed×Drug	$(2 - 1)(2 - 1) = 1$
Error	$(2 - 1)2(5 - 1) + (2 - 1) \cdot 2 \cdot 5 \cdot 2 = 28$
c. Total	39

The error degrees of freedom can be obtained by subtraction or by summing the degrees of freedom for Drug × Pen(Feed) and Pig(Feed, Pen, Drug).

Because pens are the experimental units for the feed-type treatment factor, the mean square for Pen(Feed) is the denominator term for testing for a feed-type main effect. Thus, the degrees of freedom are 1 and 8 for the previous F statistic.

8. The drug main effect is $\delta_1 - \delta_2 + \bar{\gamma}_{.1} - \bar{\gamma}_{.2}$, whose BLUE is $\bar{y}_{..1.} - \bar{y}_{..2.}$. We have

$$\text{Var}(\bar{y}_{..1.} - \bar{y}_{..2.}) = \text{Var}(\bar{e}_{..1.} - \bar{e}_{..2.}) = 2\sigma_e^2/20 = 0.1\sigma_e^2.$$

Thus, a t statistic for testing for a drug main effect is

$$\frac{\bar{y}_{.1..} - \bar{y}_{..2.}}{\sqrt{0.1\hat{\sigma}_e^2}} = \frac{36.9 - 35.3}{\sqrt{1.26}} \approx 1.425,$$

which implies that the F statistic is approximately $1.425^2 \approx 2.03$.

9. Drug is the split-plot factor in this experiment. The Error line of the ANOVA table provides the denominator mean square for the F test of split-plot main effects. Thus, the degrees of freedom are 1 and 28 for the previous F statistic.

10. There is interaction between feed type and drug if and only if the effect of drug when feed type is 1 is different than the effect of drug when feed type is 2. Thus, we need to test whether

$$\begin{aligned} &\{(\mu + \phi_1 + \delta_1 + \gamma_{11}) - (\mu + \phi_1 + \delta_2 + \gamma_{12})\} - \{(\mu + \phi_2 + \delta_1 + \gamma_{21}) - (\mu + \phi_2 + \delta_2 + \gamma_{22})\} \\ = \;& \gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22} = 0. \end{aligned}$$

11. The BLUE of $\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$ is $\bar{y}_{1.1.} - \bar{y}_{1.2.} - \bar{y}_{2.1.} + \bar{y}_{2.2.}$, and

$$\text{Var}(\bar{y}_{1.1.} - \bar{y}_{1.2.} - \bar{y}_{2.1.} + \bar{y}_{2.2.}) = \text{Var}(\bar{e}_{1.1.} - \bar{e}_{1.2.} - \bar{e}_{2.1.} + \bar{e}_{2.2.}) = 4\sigma_e^2/10 = 0.4\sigma_e^2.$$

Thus, the appropriate confidence interval is

$$\begin{aligned} 32.7 - 31.9 - 41.1 + 38.7 &\pm t_{.975,28}\sqrt{0.4 \cdot 12.6} \\ -1.6 &\pm 2.048\sqrt{5.04} \\ -1.6 &\pm 4.6 \Longleftrightarrow (-6.2, 3.0). \end{aligned}$$

12. The model has $2 \cdot 2 \cdot 21 = 84$ mean parameters and $1 + 21 \cdot (21+1)/2 = 232$ variance parameters, for a total of $84 + 232 = 316$ parameters. The dimension of the model's parameter space is 316.

13.

$$\begin{aligned} \text{Cov}(y_{ijkl1}, y_{ijkl7}) &= \text{Cov}(p_{ij} + e_{ijkl1}, p_{ij} + e_{ijkl7}) \\ &= \text{Cov}(p_{ij}, p_{ij}) + \text{Cov}(e_{ijkl1}, e_{ijkl7}) + \text{Cov}(p_{ij}, e_{ijkl7}) + \text{Cov}(e_{ijkl1}, p_{ij}) \\ &= \text{Var}(p_{ij}) + \sigma_e^2\rho^6 + 0 + 0 \\ &= \sigma_p^2 + \sigma_e^2\rho^6 \end{aligned}$$

14. The BLUE of $\mu_{111} - \mu_{117}$ is $\bar{y}_{1.1.1} - \bar{y}_{1.1.7}$. We have

$$\begin{aligned} \text{Var}(\bar{y}_{1.1.1} - \bar{y}_{1.1.7}) &= \text{Var}(\bar{e}_{1.1.1} - \bar{e}_{1.1.7}) = \text{Var}\left[\frac{1}{10}\sum_{j=1}^{5}\sum_{l=1}^{2}(e_{1k1l1} - e_{1k1l7})\right] \\ &= 0.1\text{Var}(e_{11111} - e_{11117}) = 0.1[\sigma_e^2 + \sigma_e^2 - 2\text{Cov}(e_{11111}, e_{11117})] \\ &= 0.2(\sigma_e^2 - \sigma_e^2\rho^6) = 0.2\sigma_e^2(1 - \rho^6). \end{aligned}$$

Thus, a t statistic for testing $H_0 : \mu_{111} = \mu_{117}$ is

$$\frac{\bar{y}_{1.1.1} - \bar{y}_{1.1.7}}{\sqrt{0.2\hat{\sigma}_e^2(1 - \hat{\rho}^6)}} = \frac{1.99 - 1.69}{\sqrt{0.2 \cdot 0.8 \cdot (1 - 0.9^6)}} \approx 3.70$$

15. The AIC should be greater than 1061.8 for model (3B) because smaller values of AIC are preferred.

16. $1061.8 = -2 \times (-527.9) + 2 \times 3$, where 3 is the number of variance parameters in model (3A) (i.e., σ_p^2, σ_e^2, ρ). Only the number of variance parameters (rather than the total number of parameters) is used because the REML likelihood involves only the variance parameters.

17. The residual log likelihood for model (3) must be larger than -527.9 because model (3) is a more flexible model that involves the same mean parameters and far more variance parameters than model (3A). Model (3A) is a special case of model (3), so the maximized residual log likelihood cannot be greater for model (3A) than for model (3).

18. The likelihood ratio statistic is $-2\{(-490)-(-435)\}=110$. The degrees of freedom is equal to the difference between the dimensions of the parameters spaces of the models, which is $87-11=76$. A chi-squared distribution with 76 degrees of freedom has mean 76 and standard deviation $\sqrt{2\times 76}\approx 12.3$. Thus, the likelihood ratio statistic of 110 is about 2.76 standard deviations above the mean under the null, so there is reason to question the fit of model (3AL). More formally, the p-value for the likelihood ratio test is approximately 0.007. Thus, we would reject the simpler null model (3AL) in favor of the alternative model (3A). It does not seem that model (3AL) provides an adequate fit to the data relative to model (3A).

19. The MLE of π is $\hat{\pi}=150/500=0.3$. Thus, a Wald interval for π with coverage probability approximately equal to 0.95 is

$$0.3\pm 1.96\sqrt{0.3\times(1-0.3)/500}\iff(0.26,0.34).$$

20. The Delta method can be used to find an interval with coverage probability approximately equal to 0.95. The MLE of

$$\lambda=\pi+(1-\pi)\{1-G(0.75;\theta)\}=1-G(0.75;\theta)+\pi G(0.75;\theta)$$

is

$$\hat{\lambda}=\hat{\pi}+(1-\hat{\pi})\{1-G(0.75;\hat{\theta})\}=0.3+0.7\{1-G(0.75;4)\}=1-0.7G(0.75;4).$$

The inverse of the observed Fisher information matrix is diagonal with diagonal elements

$$\hat{v}_1\equiv\hat{\pi}(1-\hat{\pi})/500 \text{ and } \hat{v}_2\equiv\left[\sum_{i=1}^{350}\frac{\{G^{(1,1)}(y_i;\hat{\theta})\}^2-G^{(1,0)}(y_i;\hat{\theta})G^{(1,2)}(y_i;\hat{\theta})}{\{G^{(1,0)}(y_i;\hat{\theta})\}^2}\right]^{-1}.$$

The partial derivative of λ with respect to π, evaluated at the MLE of (π,θ), is $d_1\equiv G(0.75;4)$. The partial derivative of λ with respect to θ, evaluated at the MLE of (π,θ), is $d_2\equiv-0.7G^{(0,1)}(0.75;4)$. Thus, an approximation to the variance of $\hat{\lambda}$ is

$$\hat{v}=d_1^2\hat{v}_1+d_2^2\hat{v}_2,$$

and our interval for λ with coverage probability approximately equal to 0.95 is

$$\hat{\lambda}\pm 1.96\sqrt{\hat{v}}.$$

These are a sketch of the answers hoped for. Other possibilities might exist for some of the questions that would be entirely adequate if they are both technically correct and logically consistent.

Question 1. To find the expected value and variance of a random variable Z, having density given in expression (1) in the exam question, note that

$$\begin{aligned}
\int_{-1}^{0} z f_n(z|\alpha,\beta)\,dz &= -\frac{\alpha}{\alpha+\beta} = -\mu \\
\int_{0}^{1} z f_p(z|\alpha,\beta)\,dz &= \frac{\alpha}{\alpha+\beta} = \mu \\
\int_{-1}^{0} z^2 f_n(z|\alpha,\beta)\,dz &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = \phi\mu(\alpha+1) = \phi\mu(1-\mu)+\mu^2 \\
\int_{0}^{1} z^2 f_p(z|\alpha,\beta)\,dz &= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = \phi\mu(\alpha+1) = \phi\mu(1-\mu)+\mu^2,
\end{aligned}$$

the final step in the last two lines following from substitution of $\alpha = [(1/\phi)-1]\mu$. The expected value of a random variable Z that has the density (1) is then,

$$\begin{aligned}
E(Z) &= \gamma\int_{-1}^{0} z f_n(z|\alpha,\beta)\,dz + (1-\gamma)\int_{0}^{1} z f_p(z|\alpha,\beta)\,dz \\
&= -\gamma\mu + (1-\gamma)\mu = \mu(1-2\gamma),
\end{aligned}$$

The second moment is,

$$\begin{aligned}
E(Z^2) &= \gamma\int_{-1}^{0} z^2 f_n(z|\alpha,\beta)\,dz + (1-\gamma)\int_{0}^{1} z^2 f_n(z|\alpha,\beta)\,dz \\
&= \gamma\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} + (1-\gamma)\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \\
&= \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} = \phi\mu(\alpha+1) = \phi\mu(1-\mu)+\mu^2.
\end{aligned}$$

The variance of Z may then be found as,

$$var(Z) = E(Z^2) - [E(Z)]^2 = \phi\mu(1-\mu) + \mu^2 - \mu^2(1-2\gamma)^2.$$

Finally, if $Y = \lambda Z$, then

$$\begin{aligned}
E(Y) &= \lambda\mu(1-2\gamma), \\
var(Y) &= \lambda^2\phi\mu(1-\mu) + \lambda^2\mu^2 - \lambda^2\mu^2(1-2\gamma)^2 \\
&= \lambda^2\phi\mu(1-\mu) + \lambda^2 4\mu\gamma(1-\mu\gamma).
\end{aligned}$$

It is not possible, then, to write the variance $var(Y)$ as proportional to some simple function of the mean $E(Y)$.

Question 2. Using the expected value $E(Y) = \lambda\mu(1-2\gamma)$ it can be seen that, for a fixed γ, $E(Y)$ is either strictly positive (if $\gamma < 0.5$), strictly negative (if $\gamma > 0.5$), or identically zero (if $\gamma = 0.5$). This is because $0 < \mu < 1$. For a fixed μ, however, $E(Y)$ varies from $-\lambda\mu$ to $\lambda\mu$ as γ changes. Thus, covariate information should be incorporated into a model for γ that vary over states as γ_i.

Question 3. For a random sample $Y_1, \ldots, Y_n$ from the density given in expression (2) of the exam question the log likelihood is $\ell = \sum \ell_i$ where,

$$\begin{aligned} \ell_i &= [\log(K) + \log(1-\gamma) - (\alpha+\beta-1)\log(\lambda) + (\alpha-1)\log(|y_i|) + (\beta-1)\log(\lambda-|y_i|)]I(y_i < 0) \\ &+ [\log(K) + \log(\gamma) - (\alpha+\beta-1)\log(\lambda) + (\alpha-1)\log(y_i) + (\beta-1)\log(\lambda-y_i)]I(y_i > 0), \quad (1) \end{aligned}$$

where $I(A)$ is the indicator function that assumes a value of 1 if A is true and 0 otherwise, and

$$K = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}.$$

Question 4. With the log likelihood given in Question 3, let $N_n = \sum I(y_i < 0)$ and $N_p = \sum I(y_i > 0)$ be the numbers of negative and positive responses. Then we have that

$$\frac{\partial}{\partial\gamma}\ell = -\frac{N_n}{1-\gamma} + \frac{N_p}{\gamma},$$

and,

$$\hat{\gamma} = \frac{N_p}{N_p + N_n}$$

Question 5. From (1) in Question 3, it may be seen that the log likeilhood can be written as,

$$\ell_i = \sum_{i=1}^{n} [\log(1-\gamma)I(y_i < 0) + \log(\gamma)I(y_i > 0)] + Q(\alpha, \beta, \lambda|\boldsymbol{y}),$$

Q can, in turn, be written as the log likelihood of a random sample from a scaled beta distribution, evaluated at the absolute values of the y_i; $i = 1, \ldots, n$

$$Q = \sum_{i=1}^{n}[\log(K) - (\alpha+\beta-1)\log(\lambda) + (\alpha-1)\log(|y_i|) + (\beta-1)\log(\lambda-|y_i|)]I(y_i < 0)$$

$$
\begin{aligned}
&+ \sum_{i=1}^{n}[\log(K) - (\alpha+\beta-1)\log(\lambda) + (\alpha-1)\log(y_i) + (\beta-1)\log(\lambda - y_i)]I(y_i > 0) \\
&= \sum_{i=1}^{n}[\log(K) - (\alpha+\beta-1)\log(\lambda) + (\alpha-1)\log(|y_i|) + (\beta-1)\log(\lambda - |y_i|)].
\end{aligned}
$$

Question 6. Following the hint, we can write the likelihood for a random sample of size n from a uniform distribution on $(0, \lambda)$ as,

$$L(\theta|\boldsymbol{y}) = \frac{1}{\lambda^n}$$

If we let $\hat{\lambda} = y_{[n]}$ the maximized likelihood is greater than 0.

For a random sample from a scaled beta distribution (with $\beta \neq 1$), the likelihood is

$$L(\alpha, \beta, \lambda) = \left[K\frac{1}{\lambda^{\alpha+\beta-1}}\right]^n \prod_{i=1}^{n} y_i^{\alpha-1} \prod_{i=1}^{n}(\lambda - y_i)^{\beta-1}.$$

This likelihood, evaluated at $\lambda = y_{[n]}$ is 0 because of the one term in which $y_i = y_{[n]} = \lambda$. What we might call the data-dependent parameter space then could be specified as $y_{[n]} \leq \lambda < \infty$ for the uniform problem (also the scaled beta problem if $\beta = 1$), but must be taken as $y_{[n]} < \lambda < \infty$ for the scaled beta with $\beta \neq 1$.

An equivelant way to make the same point is to notice that the support of a uniform distribution may be taken as $y \in (0, \lambda]$, while the support of a scaled beta with $\beta \neq 1$ must be $y \in (0, \lambda)$.

Question 7. The derivatives with respect to the regression parameters of the log likelihood written as in the answer to either Question 3 or Question 4 are,

$$
\begin{aligned}
\frac{\partial \ell}{\partial \psi_0} &= \sum_{i=1}^{n} \frac{\partial \ell}{\partial \gamma_i}\frac{d\gamma_i}{d\eta_i}\frac{\partial \eta_i}{\partial \psi_0} \\
&= \sum_{i=1}^{n}\left[\left(\frac{-1}{1-\gamma_i}\right) I(y_i < 0) - \frac{1}{\gamma_i}I(y_i > 0)\right]\gamma_i(1-\gamma_i) \\
\frac{\partial \ell}{\partial \psi_1} &= \sum_{i=1}^{n} \frac{\partial \ell}{\partial \gamma_i}\frac{d\gamma_i}{d\eta_i}\frac{\partial \eta_i}{\partial \psi_0} \\
&= \sum_{i=1}^{n}\left[\left(\frac{-1}{1-\gamma_i}\right) I(y_i < 0) - \frac{1}{\gamma_i}I(y_i > 0)\right]\gamma_i(1-\gamma_i)x_i.
\end{aligned}
$$

These are the same score functions that we would obtain from a basic glm with binary random component and logit link. That is, we could define random variables

$$Z_i = \begin{cases} 1 & \text{if } Y_i > 0 \\ 0 & \text{if } Y_i < 0 \end{cases}$$

Assigning the Z_i indpendent binary distributons with parameters γ_i for $i = 1, \ldots, n$ and modeling γ_i as in the model of the question leads to the score equations in (2).

Question 8. (a) Inference About ψ_0 and ψ_1.

(i) As described in Question 7, maximum likelihood estimates of ψ_0 and ψ_1 may be found independently of the other parameters. This also implies that the full information matrix (assuming it exists) for all of the parameters is block diagonal with one block consisting of the 2×2 matrix for the regression parameters and the other block containing values for the other three parameters. Computation of intervals based on asymptotic normality of mles seems applicable here.

(ii) Because of the separation of the likelihood aleardy noted, profile likelihoods for any of the other parameters result immediately in maximum likelihood estimation of ψ_0 and ψ_1. There does not seem to be any motivation whatsoever to profile either of the regression parameters, so the use of profiling methods does not seem appropriate here.

(iii) One could certainly use parametric bootstrap to compute intervals for ψ_0 and ψ_1. Asymptotic normality motivates the use of a compartison function that is either a simple difference or a studentized version of the difference.

(b) Inference About α and β.

(i) One could make use of Wald Theory conditional on a given value of λ, but the estimtaes of these parameters that result from the procedure outlined previously are not marginal maximum likelihood estimtaes, so one could also question this approach.

(ii) There apears little motivation to proflie either α or β, and it is questionable whether the profile likelihood in either of these parameters exists. That is, it is not clear that a likelihood with fixed α or fixed β has a maximum in λ other than the boundary of the data-dependent parameter space. It would be highly questionable whether this approach would provide any useful resulsts.

(iii) Parametric bootstrap would be a viable possibility for developing intervals for α and β. Comparison functions may not be crystal clear for these parameters. One possibility would be to use a mean-value parameterization with a difference comparison function for the location parameter (mean) and perhaps a ratio for the dispersion parameter. It would be interesting to compare this to the naive use of difference comparison functions for both α and β.

(c) Inference About λ.

(i) Inference based on asymptotic properties of maximum likelihood estimators is clear not appropriate for inference about λ. We do not have a regular problem with respect to this parameter, and we don't intend to find maximum likelihood estimates, even if they might exist for particular data sets.

(ii) Basing inference about λ on profile likelihood also is not appropriate. The difficulty with this parameter was originally identified because the profile likelihood in λ did not appear to have a maximum in the data-dependent parameter space of $\lambda \in (y_{[n]}, \infty)$.

(iii) A parametric bootstrap approach would also be problematic to apply to λ. The probability that the estimated value of λ for a bootstrap sample is less than the maximum observed value in the actual data set is quite high and would be expected to occur for quite a few of bootstrap samples. There does not appear to be an obvious solution to this difficulty.

Question 9. Using a mean value parameterization with

$$\begin{aligned} \mu &= \frac{\alpha}{\alpha+\beta} \\ \phi &= \frac{1}{\alpha+\beta+1} \end{aligned}$$

makes the choice of uniform prior distributions for each parameter individually a natural choice.

Question 10. Full conditional posteriors may be developed from the expression for the joint posterior, to which each full conditional is proportional,

$$p(\psi_0,\psi_1,\alpha,\beta,\lambda|\boldsymbol{y}) \propto \pi(\psi_0,\psi_1,\alpha,\beta,\lambda) f_Y(\boldsymbol{y}|\psi_0,\psi_1,\alpha,\beta,\lambda).$$

We will assume the joint prior has been specified in product form and will use the notation given in the hint to the question. For each full conditional we then keep only the relevant terms from the joint data density and obtain,

$$\begin{aligned} p(\psi_0|\cdot) &\propto \pi(\psi_0) \prod_{y_i\in C_n} (1-\gamma_i) \prod_{y_i\in C_p} \gamma_i, \\ p(\psi_1|\cdot) &\propto \pi(\psi_1) \prod_{y_i\in C_n} (1-\gamma_i) \prod_{y_i\in C_p} \gamma_i, \\ p(\alpha|\cdot) &\propto \pi(\alpha) \left(\frac{K}{\lambda^{\alpha+\beta-1}}\right)^n \prod_{i=1}^n |y_i|^{\alpha-1}, \\ p(\beta|\cdot) &\propto \pi(\beta) \left(\frac{K}{\lambda^{\alpha+\beta-1}}\right)^n \prod_{i=1}^n (\lambda-|y_i|)^{\beta-1}, \\ p(\lambda|\cdot) &\propto \pi(\lambda) \left(\frac{1}{\lambda^{\alpha+\beta-1}}\right)^n \prod_{i=1}^n (\lambda-|y_i|)^{\beta-1}. \end{aligned}$$