Part I

Protein contents of three types of forage crops were studied in a randomized block experiment using 6 blocks. Forage crops provide food for animals such as sheep, horses, and beef and dairy cattle. The experiment was performed at six different locations in different parts of Iowa. Differences across locations in soil types, rainfall, and other environmental conditions could have significant effects on the protein contents of the forage crops. In this study, each forage crop was planted in one of three plots available at each location. (The forage crops were randomly assigned to the plots within each location.) Each plot was divided into four subplots that were harvested on four different dates: 10 weeks, 14 weeks, 18 weeks, and 22 weeks after planting. The harvest dates were randomly assigned to the four subplots, using a separate random assignment within each plot. Two samples of plant material (replicates) were taken from the plant material harvested from each subplot. The protein content of the plant material was determined for each of the replicates.

The layout of the plots and subplots for one location is displayed below in Figure 1. The three types of forage crops are designated as crop 1, crop 2, and crop 3.

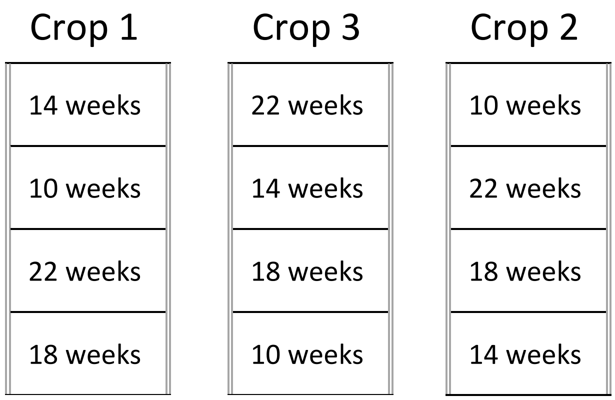


Figure 1. Forage crops planting positions and harvest dates at one location

Table 1 shows sums of squares computed from the protein content data from this experiment and associated degrees of freedom. You may find it necessary to combine sums of squares or partition sums of squares to help answer the questions. Table 2 displays sample means for the protein content for each of the three forage crops on each of the four harvest dates, averaging across locations and replicates.

Source of Variation	Degrees of Freedom	Sum of Squares
Locations (L)	5	31,892
Crops (C)	2	12,768
L×C Interaction	10	13,372
Dates (D)	3	1,080
L×D Interaction	15	864
C×D Interaction	6	5280
L×C×D Interaction	30	2612
Replicates (R)	1	15
L×R Interaction	5	188
C×R Interaction	2	55
L×C×R Interaction	10	190
D ×R Interaction	3	50
L×D×R Interaction	15	262
C×D×R Interaction	6	244
L×C×D×R Interaction	30	410

Table 1: Sums of Squares for the Forage Crop Study

	Date of Harvest (Weeks after Planting)				
Crop	10	14	18	22	Mean
1	60	66	79	87	73
2	54	56	55	59	56
3	57	52	49	46	51
Mean	57	58	61	64	

Table 2: Sample Means for the Forage Crop Study

The researchers wanted to base data analysis on the model

$$Y_{ijk\ell} = \mu + \alpha_j + \delta_k + \gamma_{jk} + \beta_i + \eta_{ij} + \psi_{ijk} + \epsilon_{ijk\ell} \tag{1}$$

where

$Y_{ijk\ell}$ is the observed protein content of the ℓ-th replicate of the plant material harvested on the k-th date from the plot assigned to the j-th crop within the i-th location.

μ is a constant.

α_j is a parameter associated with the j-th crop.

δ_k is a parameter associated with the k-th harvest date.

γ_{jk} is a parameter associated with the mean protein content when the j-th crop is harvested on the k-th date.

β_i is a random location (block) effect and it is assumed that the random location effects are i.i.d. with $\beta_i \sim N(0, \sigma^2_\beta)$.

η_{ij} is a random plot effect and it is assumed that the random plot effects are i.i.d. with $\eta_{ij} \sim N(0, \sigma^2_\eta)$.

ψ_{ijk} is a random effect and it is assumed that these random effects are i.i.d. with $\psi_{ijk} \sim N(0, \sigma^2_\psi)$.

$\epsilon_{ijk\ell}$ is a random error and it is assumed that the random errors are i.i.d. with $\epsilon_{ijk\ell} \sim N(0, \sigma^2_\epsilon)$.

It is also assumed that all random effects are jointly independent.

1. Using the information presented above, construct an appropriate ANOVA table showing the values of sums of squares, mean squares, and degrees of freedom for the sources of variation in protein content measurements. Compute values of F-statistics for crop main effects, date main effects, and crop×date interactions. For each F-statistic, report the corresponding null hypothesis and alternative hypothesis, using the parameters in model (1).

2. Assuming model (1) is correct, compute values of estimates for the variance components σ^2_β, σ^2_η, σ^2_ψ, and σ^2_ϵ. Describe the method you used to obtain your estimates.

3. Assuming model (1) is correct, derive a formula for the variance of $\overline{Y}_{..4.} - \overline{Y}_{..1.}$, the difference between the sample means for the protein contents of the crops at 22 and 10 weeks after planting. These sample means are computed by averaging across the six locations, three crops, and the replicates. Show your work. Based on your formula, compute a standard error of $\overline{Y}_{..4.} - \overline{Y}_{..1.}$.

4. Assuming model (1) is correct, derive a formula for the variance of $\overline{Y}_{\cdot 14\cdot} - \overline{Y}_{\cdot 24\cdot}$, the difference between the sample means for the protein contents for forage crops 1 and 2 at the last harvest date. Show your work. Compute a standard error of $\overline{Y}_{\cdot 14\cdot} - \overline{Y}_{\cdot 24\cdot}$. Explain how you would use these results to construct a 95 percent confidence interval for $E(\overline{Y}_{\cdot 14\cdot}) - E(\overline{Y}_{\cdot 24\cdot})$. (Note: You do not need to report numerical values for the endpoints of the confidence interval; just show how to construct it.)

5. The researchers are considering a straight line model to describe the relationship between the mean protein content (averaging across crops) and the harvest date. Obtain the proportion of the main-effect sum of squares for dates associated with this straight line relationship. Is there a significant linear relationship? Is there any evidence of a nonlinear relationship?

6. A straight line model for the relationship between the mean protein content and the harvest date may be fit separately for each forage crop. Perform a test of the null hypothesis that the slopes of the lines are the same for all three crops. Are there significant differences between the slopes for the three forage crops?

7. Describe how mean protein contents of the three forage crops are affected by the date of harvest. You may use graphs or more formal inference procedures to support your conclusions.

Part II

8. Explain what is meant when a statistician says that a linear combination of parameters in a linear model is estimable.

9. Under model (1), the expected average protein content for the samples of the plant material harvested on the k-th date from plots assigned to the j-th crop is

$$E(\overline{Y}_{ij\cdot\cdot}) = \mu + \alpha_j + \delta_k + \gamma_{jk}. \tag{2}$$

Is γ_{12} estimable for the study described on the first page of this problem? Support your answer by either carefully showing that γ_{12} is estimable or carefully showing that γ_{12} is not estimable.

10. When model (1) is fit to the data from the study, the SAS, JMP and R software packages produce different sets of estimates for the parameters in the expected protein content formula presented in **problem 9**. Using your knowledge of the theory of least squares estimation for linear models, explain why this happens. Explain why the users of the SAS, JMP, and R statistical software packages should not be concerned about the disparities in the sets of parameter estimates provided by these three packages.

Part III

11. Using the model (1) notation, derive the variance-covariance matrix for

$$\overline{\mathbf{Y}}_{ij} = \begin{bmatrix} \overline{Y}_{ij1\cdot} \\ \overline{Y}_{ij2\cdot} \\ \overline{Y}_{ij3\cdot} \\ \overline{Y}_{ij4\cdot} \end{bmatrix},$$

the vector of sample means for the four harvest dates from the plot in which the j-th crop is planted at the i-th location. The sample means are ordered so that $\overline{Y}_{ij1\cdot}$ is the mean protein content for the plant material harvested 10 weeks after planting, $\overline{Y}_{ij2\cdot}$ is the mean protein content for the plant material harvested 14 weeks after planting, $\overline{Y}_{ij3\cdot}$ is the mean protein content for the plant material harvested 18 weeks after planting, and $\overline{Y}_{ij4\cdot}$ is the mean protein content for the plant material harvested 22 weeks after planting.

12. Suppose that variation in the protein content among plants increases with age for each of the forage crops. Further suppose that the correlation between protein contents of plants harvested from the same plot at two different harvest dates is weaker when time between harvest dates is longer. Suggest a better model for the covariance matrix for $\overline{\mathbf{Y}}_{ij}$ than the covariance matrix you derived in **problem 11**.

13. If possible, explain how to test the null hypothesis that the model for the covariance matrix you derived in **problem 11** is the correct covariance matrix for $\overline{\mathbf{Y}}_{ij}$ against the alternative hypothesis that the model you proposed in **problem 12** is correct. If you believe that a hypothesis test cannot be performed, describe some other way to decide which of the two covariance models is more plausible.

Part I

1. Using the information given on page 2 of the question, the following ANOVA table may be constructed for the sources of variations corresponding to the fixed and random effects in model (1).

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F Statistic
Whole Plot Effects:				
Locations (L)	5	31892	6378.4	
Crops (C)	2	12768	6384.0	4.77
Whole Plots within Locations	10	13372	1337.2	
Subplot Effects:				
Dates (D)	3	1080	360	4.66
C×D Interaction	6	5280	880	11.39
Subplots within Plots	45	3476	77.2444	
Replicates:				
Replicates within subplots	72	1414	19.6389	
Corrected Total	143	69282		

The requested sums of squares, mean squares, degrees of freedom, and F-statistics are displayed above.

The null hypothesis associated with the F-test for crop effects is

$$H_0 : \alpha_1 + \frac{1}{4}\sum_{k=1}^{4}\gamma_{1k} = \alpha_2 + \frac{1}{4}\sum_{k=1}^{4}\gamma_{2k} = \alpha_3 + \frac{1}{4}\sum_{k=1}^{4}\gamma_{3k}$$

and the alternative is that at least two of the quantities constrained to be equal by the null hypothesis are not equal.

The null hypothesis associated with the F-test for date effects is

$$H_0 : \delta_1 + \frac{1}{3}\sum_{j=1}^{3}\gamma_{j1} = \delta_2 + \frac{1}{3}\sum_{j=1}^{3}\gamma_{j2} = \delta_3 + \frac{1}{3}\sum_{j=1}^{3}\gamma_{j3} = \delta_4 + \frac{1}{3}\sum_{j=1}^{3}\gamma_{j4}$$

and the alternative is that at least two of the quantities constrained to be equal by the null hypothesis are not equal.

The null hypothesis associated with the F-test for crop×date interaction effects is that all interaction contrasts are zero. This can be expressed as

$H_0 : \gamma_{jk} - \gamma_{jt} - \gamma_{sk} + \gamma_{st} = 0$ for all $j = 1, 2, 3,\ s = 1, 2, 3,\ k = 1, 2, 3, 4,$ and $t = 1, 2, 3, 4.$

The alternative is that at least one of the interaction contrasts is not zero.

2. Because it is a balanced experiment, the expectations of the mean squares associated with the random effects in model (1) have simple formulas:

$$\begin{aligned} E(MS_{error}) &= \sigma^2_\epsilon \\ E(MS_{subplots}) &= \sigma^2_\epsilon + 2\sigma^2_\psi \\ E(MS_{wholeplots}) &= \sigma^2_\epsilon + 2\sigma^2_\psi + 8\sigma^2_\eta \\ E(MS_{locations}) &= \sigma^2_\epsilon + 2\sigma^2_\psi + 8\sigma^2_\eta + 24\sigma^2_\beta \end{aligned}$$

Unbiased method-of-moments estimates are obtained by setting the respective formulas equal to the values of the mean squares from the ANOVA table. The estimates of the variance components are $\hat{\sigma}^2_\epsilon = 19.64$, $\hat{\sigma}^2_\psi = \frac{77.2444-19.6389}{2} = 28.80$, $\hat{\sigma}^2_\eta = \frac{1337.2-77.2444}{8} = 157.49$, and $\hat{\sigma}_\beta = \frac{6378.4-1337.2}{24} = 210.88$.

3. Under model (1)

$$\begin{aligned} Var(\overline{Y}_{..4.} &- \overline{Y}_{..1.}) \\ &= Var\left(\frac{1}{18}\sum_{i=1}^{6}\sum_{j=1}^{3}\psi_{ij4} + \frac{1}{36}\sum_{i=1}^{6}\sum_{j=1}^{3}\sum_{\ell=1}^{2}\epsilon_{ij4\ell} - \frac{1}{18}\sum_{i=1}^{6}\sum_{j=1}^{3}\psi_{ij1} - \frac{1}{36}\sum_{i=1}^{6}\sum_{j=1}^{3}\sum_{\ell=1}^{2}\epsilon_{ij1\ell}\right) \\ &= \frac{\sigma^2_\epsilon + 2\sigma^2_\psi}{18} \end{aligned}$$

Because the expectation of the mean square for subplot variation within plots is $\sigma^2_\epsilon + 2\sigma^2_\psi$, this mean square provides an unbiased estimator, and the standard error for $\overline{Y}_{..4.} - \overline{Y}_{..1.}$ is

$$S_{\overline{Y}_{..4.} - \overline{Y}_{..1.}} = \sqrt{\frac{77.244}{18}} = 2.07.$$

4. Under model (1)

$$\begin{aligned} Var\left(\overline{Y}_{\cdot 14\cdot} - \overline{Y}_{\cdot 24\cdot}\right) &= Var\left(\frac{1}{6}\sum_{i=1}^{6}\eta_{i1} + \frac{1}{6}\sum_{i=1}^{6}\psi_{i14} + \frac{1}{12}\sum_{i=1}^{6}\sum_{\ell=1}^{2}\epsilon_{i14\ell}\right. \\ &\qquad \left. - \frac{1}{6}\sum_{i=1}^{6}\eta_{i4} - \frac{1}{6}\sum_{i=1}^{6}\psi_{i44} - \frac{1}{12}\sum_{i=1}^{6}\sum_{\ell=1}^{2}\epsilon_{i44\ell}\right) \\ &= \frac{\sigma^2_\epsilon + 2\sigma^2_\eta + 2\sigma^2_\psi}{6} \\ &= \frac{E(MS_{Whole\ plots}) + 3E(MS_{Subplots})}{24} \end{aligned}$$

The standard error is

$$S_{\overline{Y}_{\cdot 14\cdot}-\overline{Y}_{\cdot 44\cdot}} = \sqrt{\frac{1337.2+(3)(77.2444)}{24}} = 8.09.$$

An approximate 95 percent confidence interval is

$$(\overline{Y}_{\cdot 14\cdot} - \overline{Y}_{\cdot 44\cdot}) \pm t_{(df,0.975)} S_{\overline{Y}_{\cdot 14\cdot}-\overline{Y}_{\cdot 44\cdot}}$$

where the degrees of freedom are obtained from the Cochran-Satterthwaite formula

$$df = \frac{(MS_{Whole\ plots} + (3)MS_{subplots})^2}{\frac{(1)^2 MS^2_{Whole\ plots}}{10} + \frac{(3)^2 MS^2_{Subplots}}{45}} = 13.77$$

5. The value of the linear contrast is $c_{linear} = (-3)(57) + (-1)(58) + (1)(61) + (3)(64) = 24.$ The value of the sum of squares for this contrast is

$$SS_{linear} = \frac{c^2_{linear}}{\frac{(-3)^2}{36} + \frac{(-1)^2}{36} + \frac{(1)^2}{36} + \frac{(3)^2}{36}} = 1036.8.$$

This is 96 percent of the sum of squares representing the variation in the mean protein content across harvest dates.

An F-statistic is

$$F = \frac{SS_{linear}}{MS_{subplots}} = 13.412 \qquad \text{on } (1,45) \text{ df.}$$

This F-value is larger than 7.22, the 99-th percentile of the central F-distribution with (1,45) df. Consequently, the relationship between the mean protein content (averaging across crops) and harvest date has a significant linear component.

The remaining portion of SS_{dates} is

$$SS_{remainder} = SS_{dates} - SS_{linear} = 1080 - 1036.8 = 43.2 \qquad \text{on 2 df.}$$

Because this is much smaller than $MS_{subplots} = 77.244$, there are no obvious non-linear trends in the mean protein content across the harvest dates. Alternatively, this part of the response could be based on sums of squares for orthogonal quadratic and cubic contrasts, $SS_{quadratic} = 36$ and $SS_{cubic} = 7.2$, respectively, to show that quadratic and cubic trends are not significant.

6. An F-test may be used to check for significant differences in linear trends for the mean protein content across harvest dates for the three forage crops. Compute sums of squares for two orthogonal interaction contrasts. There are many ways to pick appropriate contrasts, but one pair is

$$c_{(A-B)\times Dates:linear} = \left[-3\overline{Y}_{\cdot 11\cdot} - \overline{Y}_{\cdot 12\cdot} + \overline{Y}_{\cdot 13\cdot} + 3\overline{Y}_{\cdot 14\cdot}\right] - \left[-3\overline{Y}_{\cdot 21\cdot} - \overline{Y}_{\cdot 22\cdot} + \overline{Y}_{\cdot 23\cdot} + 3\overline{Y}_{\cdot 24\cdot}\right] = 80.$$

and

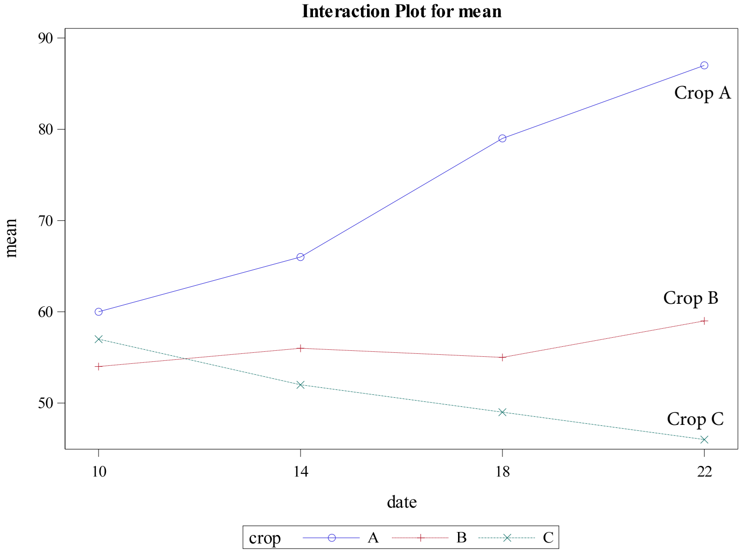
$$\begin{aligned} c_{(A+B-2C)\times Dates:linear} &= \left[-3\overline{Y}_{\cdot 11\cdot} - \overline{Y}_{\cdot 12\cdot} + \overline{Y}_{\cdot 13\cdot} + 3\overline{Y}_{\cdot 14\cdot}\right] + \left[-3\overline{Y}_{\cdot 21\cdot} - \overline{Y}_{\cdot 22\cdot} + \overline{Y}_{\cdot 23\cdot} + 3\overline{Y}_{\cdot 24\cdot}\right] \\ &\qquad -2\left[-3\overline{Y}_{\cdot 21\cdot} - \overline{Y}_{\cdot 22\cdot} + \overline{Y}_{\cdot 23\cdot} + 3\overline{Y}_{\cdot 24\cdot}\right] \\ &= 180. \end{aligned}$$

Compute

$$\begin{aligned} F &= \frac{\left[SS_{(A-B)\times Dates:linear} + SS_{(A+B-2C)\times Dates:linear}\right]/2}{MS_{subplots}} \\ &= \frac{\left[\frac{(80)^2}{\frac{40}{12}} + \frac{(180)^2}{\frac{120}{12}}\right]/2}{77.2444} \\ &= 33.4 \end{aligned}$$

on (2,45) df. Because this value is much larger than $F_{(2,45),0.99} = 5.11$, there is strong evidence that the slopes of the linear relationships between the harvest date and the mean protein content are not the same for all three forage crops.

7. Note that $SS_{(A-B)\times Dates:linear} + SS_{(A+B-2C)\times Dates:linear} = 5160$ is 97.73 percent of the interaction sum of squares between crops and harvest dates. The differences between forage crops in linear trends for the mean protein content across harvest dates are essentially all of this interaction. A profile plot, such as the one shown below, can be used to graphically illustrate the differences in the trends.



This plot shows a linear increasing trend in the mean protein content between 10 and 22 weeks after planting for forage crop 1, a less dramatic decreasing trend for forage crop 3, and relatively little change in mean protein content for forage crop 2.

Part II

8. A linear combination of parameters in a linear model is estimable if it is equal to the expectation of a linear combination of responses for cases in the study.

9. Suppose γ_{12} is estimable. Then there would be non-random coefficients $c_{11}, c_{12}, \ldots, c_{34}$ such that

$$\begin{aligned}
\gamma_{12} &= \sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk}E(Y_{jk}) \\
&= \sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk}\left(\mu + \alpha_j + \delta_k + \gamma_{jk}\right) \\
&= \mu\sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk} + \sum_{k=1}^{3}\sum_{j=1}^{4} c_{jk}\alpha_j + \sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk}\delta_k + \sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk}\gamma_{jk}
\end{aligned}$$

for any set of values for the parameters. Elimination of μ requires that $0 = \sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk}$. Elimination of the other interaction parameters requires that $c_{jk} = 0$ for all (j, k) except $(1, 2)$, and retention of γ_{12} requires that $c_{12} = 1$, but this contradicts the requirement that $0 = \sum_{j=1}^{4}\sum_{k=1}^{3} c_{jk}$. Consequently, γ_{12} is not estimable.

10. Model (1) can be expressed in matrix form as $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \epsilon$, where $\mathbf{Y}$ is a 144×1 vector of observations, ϵ is a 144×1 vector of random errors, $\mathbf{u}$ is a 120×1 vector of random effects, $\mathbf{Z}$ is the corresponding 144×120 model matrix for the random effects, β is a 20×1 vector of fixed effects, and $\mathbf{X}$ is the corresponding 144×20 model matrix for fixed effects. A least squares estimator $\mathbf{b}$ for β minimizes $(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})$ and is a solution to the normal equations $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$. Because $\mathbf{X}$ is a matrix with column rank equal to 12, $\mathbf{X}$ is not of full column rank, and there is not a unique solution to the normal equations. In fact, there are an infinite number of solutions to the normal equations and 20-12=8 linearly independent constraints must be placed on the elements of $\mathbf{b}$ to obtain a particular solution. Different software packages use different sets of constraints to obtain different solutions.

This is not a concern because the estimate of any estimable combination of parameters is the same for any solution to the estimating equations. The expected value of the protein content is estimable for each crop and harvest date combination, so the estimates of the expected protein contents for the twelve combinations of crops and harvest dates are the same for any solution to the estimating equations.

Part III

11. The covariance matrix for $\overline{Y}_{ij}$ is

$$\begin{bmatrix} \sigma_\beta^2+\sigma_\eta^2+\sigma_\psi^2+0.5\sigma_\epsilon^2 & \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2 \\ \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2+\sigma_\psi^2+0.5\sigma_\epsilon^2 & \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2 \\ \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2+\sigma_\psi^2+0.5\sigma_\epsilon^2 & \sigma_\beta^2+\sigma_\eta^2 \\ \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2 & \sigma_\beta^2+\sigma_\eta^2+\sigma_\psi^2+0.5\sigma_\epsilon^2 \end{bmatrix}$$

12. The covariance matrix from **problem 11** implies the same correlation between any two subplots (harvest dates) within a plot. There are a number of options for imposing correlations such that protein content measurements at harvest dates closer in time have higher correlation than at harvest dates farther apart in time. One possibility is an autoregressive model in which the correlation between observations taken $4k$ weeks apart is ρ^k. It may be that there is less variation in protein content among younger plants than among older plants. Then, different variances, σ_{10}^2, σ_{14}^2, σ_{18}^2, σ_{22}^2, could be used at harvest dates of 10, 14, 18 and 22 weeks, respectively. The resulting covariance model for for $\overline{\mathbf{Y}}_{ij}$ is

$$R = \begin{bmatrix} \sigma_{10}^2 & \rho\sigma_{10}\sigma_{14} & \rho\sigma_{10}\sigma_{18} & \rho\sigma_{10}\sigma_{22} \\ \rho\sigma_{10}\sigma_{14} & \sigma_{14}^2 & \rho\sigma_{14}\sigma_{18} & \rho\sigma_{14}\sigma_{22} \\ \rho\sigma_{10}\sigma_{18} & \rho\sigma_{14}\sigma_{18} & \sigma_{18}^2 & \rho\sigma_{18}\sigma_{22} \\ \rho\sigma_{10}\sigma_{22} & \rho\sigma_{14}\sigma_{22} & \rho\sigma_{18}\sigma_{22} & \sigma_{22}^2 \end{bmatrix}$$

13. If the model in **problem 11** is nested in the model proposed in **problem 12**, a likelihood ratio test can be performed. For non-nested models some type of penalized likelihood or AIC method could be proposed. A Bayesian approach could also be taken.

Part I

Amazon Mechanical Turk allows employers to hire workers who perform tasks online. One type of worker, known as a Mechanical Turk Master worker, has been certified for high-quality work based on past performance. Researchers conducted a study to compare the performance of 25 Mechanical Turk Master workers with that of 25 regular Mechanical Turk workers on an image processing task. Each of the 50 workers independently processed the same set of 40 images. Based on the quality of image processing, a score was assigned to each worker for each image, with higher scores indicative of better quality. Let y_{ijk} be the score for image i and worker j in group k, where $i = 1, \ldots, 40$, $j = 1, \ldots, 25$, and $k = 1$ for Mechanical Turk Master workers and $k = 2$ for regular Mechanical Turk workers. For $i = 1, \ldots, 40$, $j = 1, \ldots, 25$, and $k = 1, 2$, consider the model

$$y_{ijk} = \mu_k + a_i + b_{jk} + e_{ijk}, \tag{1}$$

where μ_1 and μ_2 are unknown parameters and the remaining terms are independent, normally distributed, mean-zero random variables with $\mathrm{Var}(a_i) = \sigma_a^2$ for all i, $\mathrm{Var}(b_{jk}) = \sigma_b^2$ for all j and k, and $\mathrm{Var}(e_{ijk}) = \sigma_e^2$ for all i, j, and k. Model (1) was fit to the data. The values of the Best Linear Unbiased Estimators (BLUEs) of μ_1 and μ_2 are

$$\hat{\mu}_1 = 7.04 \quad \text{and} \quad \hat{\mu}_2 = 6.67,$$

and the values of the REML estimators of the variance components are

$$\hat{\sigma}_a^2 = 0.94, \quad \hat{\sigma}_b^2 = 1.05, \quad \text{and} \quad \hat{\sigma}_e^2 = 0.25.$$

Because of the balanced design used in the Amazon Mechanical Turk study, the BLUEs and REML estimates provided above can be obtained with relatively simple calculations. In particular, the BLUEs of μ_1 and μ_2 are equal to the Ordinary Least Squares Estimators (OLSEs) that can be determined by fitting a simplified version of model (1) that excludes the a_i and b_{jk} random effects. Furthermore, the REML estimates are equal to the values of the unbiased estimators of the variance components obtained by forming appropriate linear combinations of the mean squares in an ANOVA table for this dataset.

1. Model (1) is a special case of a linear mixed-effects model that can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{y}$ is a vector of response values, $\boldsymbol{X}$ is a fixed and known matrix, $\boldsymbol{\beta}$ is a vector of fixed and unknown real-valued parameters, $\boldsymbol{Z}$ is a fixed and known matrix, $\boldsymbol{u}$ is a vector of random effects, and $\boldsymbol{e}$ is a vector of random errors, with

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix} \right),$$

where $\boldsymbol{G}$ and $\boldsymbol{R}$ are variance-covariance matrices that may depend on unknown, fixed model parameters. Provide specific expressions for $\boldsymbol{G}$ and $\boldsymbol{R}$ in the special case of model (1).

2. For the general model given in (2), explain in a few sentences how REML estimates of the variance-covariance matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ are obtained.

3. Consider again the general model given in (2). Suppose $\boldsymbol{C}$ is a known matrix such that $\boldsymbol{C\beta}$ is estimable. Provide a general expression for the BLUE of $\boldsymbol{C\beta}$.

4. In the ANOVA table for the Amazon Mechanical Turk dataset, the sum of squares for workers nested within groups can be written as

$$40\sum_{j=1}^{25}\sum_{k=1}^{2}(\bar{y}_{\cdot jk}-\bar{y}_{\cdot\cdot k})^2,\text{ where }\bar{y}_{\cdot jk}=\frac{1}{40}\sum_{i=1}^{40}y_{ijk}\text{ (for each }\ j=1,\ldots,25,k=1,2)\text{ and}$$

$$\bar{y}_{\cdot\cdot k}=\frac{1}{1000}\sum_{i=1}^{40}\sum_{j=1}^{25}y_{ijk}\text{ (for each }k=1,2).$$

Compute the sum of squares for workers nested within groups.

5. Determine the value of an F statistic that can be used to test $H_0:\mu_1=\mu_2$.

6. The F statistic computed in problem **5** has a noncentral F distribution. State the degrees of freedom for this F distribution and provide an expression for the noncentrality parameter.

7. Find a standard error for $\hat{\mu}_1$.

Part II

Suppose a basketball player attempts 25 shots at a goal from various distances. The outcome of each shot is considered a success if the ball passes through the goal and a failure otherwise. For $i=1,\ldots,25$, let x_i be the distance in feet for the ith shot, and let y_i be an indicator of success for the ith shot, where $y_i=1$ if the ith shot is successful and $y_i=0$ otherwise. Results from the 25 shots are presented in Figure 1 on page 4.

Let $\Phi(\cdot)$ be the cumulative distribution function of the standard normal distribution. Consider $x_1,\ldots,x_{25}$ to be fixed and known values. As a model for the shot outcomes, suppose $y_1,\ldots,y_{25}$ are independent and that, for $i=1,\ldots,n$,

$$y_i\sim\text{Bernoulli}(\pi_i),\text{ where }\pi_i=\Phi(\beta_0+\beta_1x_i)\tag{3}$$

for some unknown real-valued parameters β_0 and β_1. Suppose model (3) is fit to the data to obtain the following maximum likelihood estimates of β_0 and β_1, as well as the following variance and covariance approximations:

$$\hat{\beta}_0=4.25,\quad\hat{\beta}_1=-0.21,\quad\widehat{\text{Var}}(\hat{\beta}_0)=3.014,\quad\widehat{\text{Var}}(\hat{\beta}_1)=0.0075,\quad\widehat{\text{Cov}}(\hat{\beta}_0,\hat{\beta}_1)=-0.148.$$

8. Provide an expression for the likelihood function.

9. Provide an estimate of the Fisher information matrix.

10. Provide a confidence interval for β_1 that has confidence level approximately equal to 0.95.

11. Estimate the probability of success for a shot attempted by the player at a distance of 20 feet from the goal.

12. Provide a standard error for the estimate in problem **11**.

13. Estimate the distance at which the player's probability of a successful shot is 2.5%.

Now suppose the player will participate in a contest. The contest involves two independent trials. Each trial is conducted as follows. A random distance will be drawn from a normal distribution with mean 19 feet and standard deviation 4 feet. The player will attempt a shot from the randomly drawn distance. The trial is considered a success if the player's shot from the randomly drawn distance is a success. Likewise, the trial is considered a failure if the player's shot from the randomly drawn distance is a failure. The player wins the contest if both trials are successes and loses the contest otherwise. Although the $N(19, 4^2)$ distribution places some small positive probability (about 1 in a million) at nonsensical negative distances, ignore this for the purposes of all subsequent problems. Simply assume that model (3) provides the success probability for a shot taken from any distance, be it positive or negative.

14. An estimate of the probability that the player will win the contest can be expressed as $g(\Phi(u))$ for some function $g(\cdot)$ and some value u. Find $g(\cdot)$ and u.

15. Now suppose that the contest is modified to be more favorable to the player. In particular, the player may – at most once during the contest – replace the distance randomly drawn for a trial with a new distance independently drawn from a normal distribution with mean 19 feet and standard deviation 4 feet. The decision to keep the originally drawn distance or to replace it with a new distance is made by the player with knowledge of the original distance but without knowledge of the new distance. For example, suppose the original distance drawn for a trial is 21.6 feet. The player must decide whether to attempt the shot from 21.6 feet or to instead attempt the shot from a new distance that will be independently drawn from a normal distribution with mean 19 feet and standard deviation 4 feet. The new distance may turn out to be less than or greater than the original distance.

Now suppose that in this modified version of the contest, the player has successfully completed the first trial and still has the option to replace the original distance drawn for the second trial. Find a threshold c_2 such that the player's estimated probability of winning the contest will be maximized if the player chooses to shoot from the original distance drawn for the second trial when that distance is less than or equal to c_2 and to replace that distance when it is greater than c_2.

16. Consider again the modified version of the contest described in problem **15**. The previous problem involved a threshold c_2 useful for a player who has already successfully completed the first trial without replacing the distance originally drawn for that trial. Now consider a player who is just starting the contest. The new goal is to determine a threshold c_1 for the first trial such that a player's estimated probability of winning the contest will be maximized if the player chooses to shoot from the original distance drawn for the first trial when that distance is less than or equal to c_1 and to replace that distance when it is greater than c_1. Because the trials are performed sequentially, the decision for the first trial must be made without knowledge of the distance drawn for the second trial. Assume that the player will utilize threshold c_2 for the second trial if the player succeeds in the first trial without replacing the original distance drawn for that trial. Because a value for c_1 may not be possible to determine without a computer, provide an expression for c_1 that can be evaluated with a computer to receive full credit for this problem.

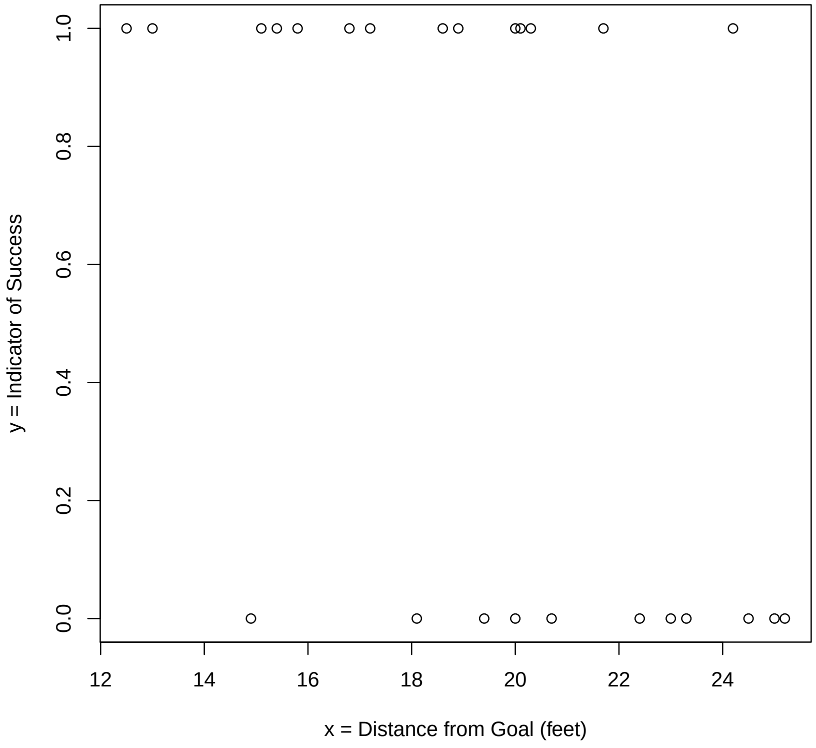


Figure 1: Scatterplot of shot outcome vs. shot distance for 25 shots.

Part I

1. Multiple answers are possible, but the most conventional is

$$\boldsymbol{G} = \begin{bmatrix} \sigma_a^2 \boldsymbol{I}_{40\times 40} & \boldsymbol{0}_{40\times 50} \\ \boldsymbol{0}_{50\times 40} & \sigma_b^2 \boldsymbol{I}_{50\times 50} \end{bmatrix} \quad \text{and} \quad \boldsymbol{R} = \sigma_e^2 \boldsymbol{I}_{2000\times 2000}.$$

2. Suppose $\boldsymbol{X}$ is an $n \times p$ matrix of rank r. Let $\boldsymbol{A}'$ be a matrix whose rows are any set of $n - r$ linearly independent rows of $\boldsymbol{I} - \boldsymbol{P_X}$. Then $\boldsymbol{w} \equiv \boldsymbol{A}'\boldsymbol{y} \sim N(\boldsymbol{0}, \boldsymbol{A}'(\boldsymbol{ZGZ}' + \boldsymbol{R})\boldsymbol{A})$. The REML estimates of $\boldsymbol{G}$ and $\boldsymbol{R}$ are obtained by replacing any parameters in $\boldsymbol{G}$ and $\boldsymbol{R}$ with their maximum likelihood estimates obtained by maximizing the multivariate normal likelihood based on the data vector $\boldsymbol{w}$.

3. If $\boldsymbol{G}$ and $\boldsymbol{R}$ are known, the BLUE is $\boldsymbol{C}(\boldsymbol{X}'\Sigma^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\Sigma^{-1}\boldsymbol{y}$, where $\Sigma = \boldsymbol{ZGZ}' + \boldsymbol{R}$. In the usual case that $\boldsymbol{G}$ and $\boldsymbol{R}$ are unknown, we approximate the BLUE by $\boldsymbol{C}(\boldsymbol{X}'\hat{\Sigma}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\hat{\Sigma}^{-1}\boldsymbol{y}$, where $\hat{\Sigma} = \boldsymbol{Z}\hat{\boldsymbol{G}}\boldsymbol{Z}' + \hat{\boldsymbol{R}}$ and $\hat{\boldsymbol{G}}$ and $\hat{\boldsymbol{R}}$ are the REML estimates of $\boldsymbol{G}$ and $\boldsymbol{R}$, respectively.

4. Because of the relationship between the REML estimates of the variance components and the ANOVA mean squares for this dataset, we can use the REML estimates provided ($\hat{\sigma}_a^2 = 0.94$, $\hat{\sigma}_b^2 = 1.05$, $\hat{\sigma}_e^2 = 0.25$) to determine the values of sums of squares. To this end, let's examine the expected value of the sum of squares for workers nested within groups:

$$\begin{aligned}
E\left(40\sum_{j=1}^{25}\sum_{k=1}^{2}(\bar{y}_{\cdot jk} - \bar{y}_{\cdot\cdot k})^2\right) &= E\left(40\sum_{j=1}^{25}\sum_{k=1}^{2}(b_{jk} - \bar{b}_{\cdot k} + \bar{e}_{\cdot jk} - \bar{e}_{\cdot\cdot k})^2\right) \\
&= 40\sum_{j=1}^{25}\sum_{k=1}^{2} E(b_{jk} - \bar{b}_{\cdot k} + \bar{e}_{\cdot jk} - \bar{e}_{\cdot\cdot k})^2 \\
&= 40\sum_{j=1}^{25}\sum_{k=1}^{2} \left\{E(b_{jk} - \bar{b}_{\cdot k})^2 + E(\bar{e}_{\cdot jk} - \bar{e}_{\cdot\cdot k})^2\right\} \\
&= 40\sum_{k=1}^{2} E\left\{\sum_{j=1}^{25}(b_{jk} - \bar{b}_{\cdot k})^2\right\} + 40\sum_{k=1}^{2} E\left\{\sum_{j=1}^{25}(\bar{e}_{\cdot jk} - \bar{e}_{\cdot\cdot k})^2\right\} \\
&= 40\sum_{k=1}^{2}\left\{24\sigma_b^2\right\} + 40\sum_{k=1}^{2}\left\{24\sigma_e^2/40\right\} \\
&= 48(40\sigma_b^2 + \sigma_e^2).
\end{aligned}$$

It follows that the sum of squares for workers nested within groups is

$$48(40 * 1.05 + 0.25) = 2028.$$

5. Because the BLUEs are OLSEs in this case, it is easy to see that $\hat{\mu}_k = \bar{y}_{\cdot\cdot k}$ for $k = 1, 2$. Thus, the BLUE of $\mu_1 - \mu_2$ is $\bar{y}_{\cdot\cdot 1} - \bar{y}_{\cdot\cdot 2}$. Note that

$$\begin{aligned}
\text{Var}(\bar{y}_{\cdot\cdot 1} - \bar{y}_{\cdot\cdot 2}) &= \text{Var}(\bar{b}_{\cdot 1} - \bar{b}_{\cdot 2} + \bar{e}_{\cdot\cdot 1} - \bar{e}_{\cdot\cdot 2}) \\
&= 2\sigma_b^2/25 + 2\sigma_e^2/1000 \\
&= \frac{2}{1000}(40\sigma_b^2 + \sigma_e^2) \\
&= \frac{2}{1000}E(\text{Mean Square for Workers within Groups}).
\end{aligned}$$

Thus, the standard error of $\bar{y}_{\cdot\cdot 1} - \bar{y}_{\cdot\cdot 2}$ is $\sqrt{0.002 * 2028/48} = \sqrt{0.0845}$, which leads to the statistic

$$t = \frac{7.04 - 6.67}{\sqrt{0.0845}} \approx 1.27$$

for testing $H_0 : \mu_1 = \mu_2$. The F statistic for testing $H_0 : \mu_1 = \mu_2$ is then

$$F = \left(\frac{7.04 - 6.67}{\sqrt{0.0845}}\right)^2 = \frac{(7.04 - 6.67)^2}{0.0845} \approx 1.62.$$

6. Because the denominator of the F statistic is proportional to the mean square for workers nested in groups, the denominator degrees of freedom is the degrees of freedom for workers nested in groups, i.e., 48. The numerator degrees of freedom is 1 because the test involves a single contrast. The noncentrality parameter is

$$\frac{(\mu_1 - \mu_2)^2}{2 * 0.002 * (40\sigma_b^2 + \sigma_e^2)} \quad \text{or} \quad \frac{(\mu_1 - \mu_2)^2}{0.002 * (40\sigma_b^2 + \sigma_e^2)},$$

depending on the convention used to define the noncentrality parameter of a noncentral F distribution. The noncentrality parameter is easily obtained by replacing estimates in the F statistic with the parameters they estimate and then dividing by two or one, again depending on the convention used to define the noncentrality parameter of a noncentral F distribution.

7. Because

$$\begin{aligned}
\text{Var}(\bar{y}_{\cdot\cdot 1}) &= \text{Var}(\bar{a}_{\cdot} + \bar{b}_{\cdot 1} + \bar{e}_{\cdot\cdot 1}) \\
&= \sigma_a^2/40 + \sigma_b^2/25 + \sigma_e^2/1000, \\
\widehat{\text{Var}}(\bar{y}_{\cdot\cdot 1}) &= \hat{\sigma}_a^2/40 + \hat{\sigma}_b^2/25 + \hat{\sigma}_e^2/1000 \\
&= 0.94/40 + 1.05/25 + 0.25/1000 = 0.06575.
\end{aligned}$$

Thus, the standard error of $\hat{\mu}_1 = \bar{y}_{\cdot\cdot 1}$ is $\sqrt{0.06575} \approx 0.2564$.

8.

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n}[\Phi(\beta_0 + \beta_1 x_i)]^{y_i}[1 - \Phi(\beta_0 + \beta_1 x_i)]^{(1-y_i)}$$

9. From the information provided in the problem statement, we have

$$\widehat{\text{Var}}\left(\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}\right) = \begin{bmatrix} 3.0140 & -0.1480 \\ -0.1480 & 0.0075 \end{bmatrix}.$$

Because

$$\hat{\boldsymbol{I}}^{-1}(\hat{\beta}_0, \hat{\beta}_1) = \widehat{\text{Var}}\left(\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}\right),$$

it follows that

$$\begin{aligned} \hat{\boldsymbol{I}}(\hat{\beta}_0, \hat{\beta}_1) &= \begin{bmatrix} 3.0140 & -0.1480 \\ -0.1480 & 0.0075 \end{bmatrix}^{-1} \\ &= \frac{1}{3.0140 * 0.0075 - 0.1480^2} \begin{bmatrix} 0.0075 & 0.1480 \\ 0.1480 & 3.0140 \end{bmatrix} \\ &= \begin{bmatrix} 10.699 & 211.127 \\ 211.127 & 4299.572 \end{bmatrix} \end{aligned}$$

10.

$$\begin{aligned} \hat{\beta}_1 \pm 1.96 * \sqrt{\widehat{\text{Var}}(\hat{\beta}_1)} &\iff -0.21 \pm 1.96 * \sqrt{0.0075} \\ &\iff -0.21 \pm 0.17 \\ &\iff (-0.38, -0.04) \end{aligned}$$

11.

$$\Phi(\hat{\beta}_0 + 20\hat{\beta}_1) = \Phi(4.25 - 4.20) = \Phi(0.05) \approx 0.52$$

12. Via the Delta Method,

$$\begin{aligned} \widehat{\text{Var}}[\Phi(\hat{\beta}_0 + 20\hat{\beta}_1)] &= [\phi(\hat{\beta}_0 + 20\hat{\beta}_1), 20\phi(\hat{\beta}_0 + 20\hat{\beta}_1)] \begin{bmatrix} 3.0140 & -0.1480 \\ -0.1480 & 0.0075 \end{bmatrix} \begin{bmatrix} \phi(\hat{\beta}_0 + 20\hat{\beta}_1) \\ 20\phi(\hat{\beta}_0 + 20\hat{\beta}_1) \end{bmatrix} \\ &= [\phi(0.05), 20\phi(0.05)] \begin{bmatrix} 3.0140 & -0.1480 \\ -0.1480 & 0.0075 \end{bmatrix} \begin{bmatrix} \phi(0.05) \\ 20\phi(0.05) \end{bmatrix} \\ &= 3.014[\phi(0.05)]^2 - 2 * 0.148 * 20[\phi(0.05)]^2 + 0.0075 * 400[\phi(0.05)]^2 \\ &= (3.014 - 2 * 0.148 * 20 + 0.0075 * 400)[\phi(0.05)]^2 \\ &= 0.094 \exp(-0.05^2)/(2\pi) \approx 0.0149. \end{aligned}$$

Thus, the standard error is $\sqrt{0.094 \exp(-0.05^2)/(2\pi)} \approx 0.122$.

13.

$$\begin{aligned} \Phi(\hat{\beta}_0 + \hat{\beta}_1 x) = 0.025 &\implies \hat{\beta}_0 + \hat{\beta}_1 x = -1.96 \\ &\implies x = (-1.96 - \hat{\beta}_0)/\hat{\beta}_1 \\ &\implies x = (-1.96 - 4.25)/(-0.21) \approx 29.57 \text{ feet} \end{aligned}$$

14. Suppose $Z \sim N(0,1)$ independent of $X \sim N(19,4^2)$. Then, with $\hat{\beta}_0 = 4.25$ and $\hat{\beta}_1 = -0.21$, the estimated probability of a successful trial may be written as

$$\begin{aligned}\Phi(\hat{\beta}_0 + \hat{\beta}_1 X) &= P(Z \le \hat{\beta}_0 + \hat{\beta}_1 X) = P(Z - \hat{\beta}_1 X \le \hat{\beta}_0)\\ &= P\left(\frac{Z - \hat{\beta}_1 X + 19\hat{\beta}_1}{\sqrt{1+16\hat{\beta}_1^2}} \le \frac{\hat{\beta}_0 + 19\hat{\beta}_1}{\sqrt{1+16\hat{\beta}_1^2}}\right)\\ &= \Phi\left(\frac{\hat{\beta}_0 + 19\hat{\beta}_1}{\sqrt{1+16\hat{\beta}_1^2}}\right) \approx \Phi(0.199) \approx 0.579.\end{aligned}$$

Thus, the estimated probability of winning the two-trial contest is $[\Phi(0.199)]^2$, which is $g(\Phi(u))$, where $g(t) = t^2$ and $u \approx 0.199$.

15. Suppose the original distance drawn for the second trial is x. Given that the first trial has already been successful, the probability of winning the contest if the player shoots from distance x is estimated to be $\Phi(\hat{\beta}_0 + \hat{\beta}_1 x)$. Given that the first trial has already been successful, the solution to the previous problem shows that the probability of winning the contest would be $\Phi(0.199)$ if the player were to reject distance x and draw a new distance. Thus, to maximize the estimated win probability, we should choose c_2 such that

$$\hat{\beta}_0 + \hat{\beta}_1 c_2 = 0.199, \text{ i.e., } c_2 = (0.199 - 4.25)/(-0.21) \approx 19.29 \text{ feet.}$$

16. Let S_1 be the event the player successfully completes the first trial without replacing the distance originally drawn for that trial. Let S_2 be the event that the player successfully completes the second trial. Let $X_2 \sim N(19,4^2)$ represent the distance originally drawn for the second trial. Then

$$\begin{aligned}P(S_2|S_1) &= P(S_2, X_2 \le c_2|S_1) + P(S_2, X_2 > c_2|S_1)\\ &= P(S_2, X_2 \le 19.29|S_1) + P(S_2, X_2 > 19.29|S_1)\\ &= \int_{-\infty}^{19.29} \Phi(4.25 - 0.21x_2)\frac{1}{4\sqrt{2\pi}}\exp\{-(x_2-19)^2/32\}dx_2\\ &\quad + P(S_2|X_2 > 19.29, S_1)P(X_2 > 19.29|S_1)\\ &\approx 0.4151 + \Phi(0.199)P(X_2 > 19.29) \text{ (numerical integral approximation)}\\ &\approx 0.4151 + \Phi(0.199)[1 - \Phi(0.0725)]\\ &\approx 0.6878.\end{aligned}$$

Now suppose the distance originally drawn for the first trial is x_1. If the player opts to shoot from distance x_1, the player's estimated probability of winning the contest is

$$\Phi(4.25 - 0.21x_1)P(S_2|S_1).$$

On the other hand, if the player rejects distance x_1 in favor of a new distance randomly drawn from $N(19, 4^2)$, the player's estimated probability of winning the contest is $[\Phi(0.199)]^2$ by the solution to problem **14**. Thus, we should choose c_1 such that

$$\Phi(4.25 - 0.21c_1)P(S_2|S_1) = [\Phi(0.199)]^2, \text{ i.e.,}$$

$$\begin{aligned} c_1 &= \left(\Phi^{-1}\left\{[\Phi(0.199)]^2/P(S_2|S_1)\right\} - 4.25\right)/(-0.21) \\ &\approx \left(\Phi^{-1}\left\{[\Phi(0.199)]^2/0.6787\right\} - 4.25\right)/(-0.21) \\ &\approx 20.4 \text{ feet.} \end{aligned}$$

Wind shear is a meteorological phenomenon that is defined as a difference in wind speed or direction over a relatively short distance in the atmosphere. Although there is horizontal wind shear, and wind shear resulting from shifts in wind directions, the classical wind shear connected with airplane accidents is vertical wind shear due to differences in wind speed at different altitudes. This vertical wind shear also has implications for the production of energy in wind farms, and can affect the function of turbines in serveral ways.

Wind shear is known to be related to temperature differences at different heights in the atmosphere, being stronger when there are greater differences in temperature over short vertical distances. This question concerns data on wind shear as it is related to temperature changes. At a weather tower near a wind farm in Iowa, the differences in wind speeds and temperatures between 250 meters and 50 meters above the earth's surface were measured several times a day for about 3 months. Note that wind turbines typically have their hubs at 80 meters above the ground. Researchers were interested in the daily maximum values of wind shear, and the associated differences in temperature. A scatterplot of daily maximum wind shear (in meters per second) versus temperature difference (in degrees celsius) is shown in Figure 1. Meteorologists and wind energy engineers are interested in modeling the way that maximum wind shear changes with changing temperature differences and, in particular, estimating the probability that maximum wind shear will be greater than 35 m/s, which is a value that can cause turbine damage if the turbine is not turned off.

Treat wind shear as responses of interest that we wish to relate to differences in temperature as a covariate. For this purpose, define random variables $Y_i;\ \ i = 1, \dots, n$ to be connected with the observed maximum wind shear on day i, and define $x_i;\ \ i = 1, \dots, n$ to be equal to the associated temperature difference when the maximum wind shear occurred.

Meteorologists provide information that maximum wind shears are believed to be distributed according to extreme value distributions, which are often used to model sample extremes (maxima and minima). The version of extreme value distribution appropriate for sample maxima has probability density function, for parameters $-\infty < \xi < \infty$ and $\theta > 0$,

$$f(y|\xi, \theta) = \frac{1}{\theta} \exp\left(-\left\{\frac{y-\xi}{\theta}\right\}\right) \exp\left[-\exp\left(-\left\{\frac{y-\xi}{\theta}\right\}\right)\right];\quad -\infty < y < \infty. \qquad (1)$$

The density (1) defines a location-scale family of distributions in which the location parameter ξ is equal to the *mode* of the distribution (rather than the expected value) and the variance is given by $(\pi^2/6)\theta^2$.

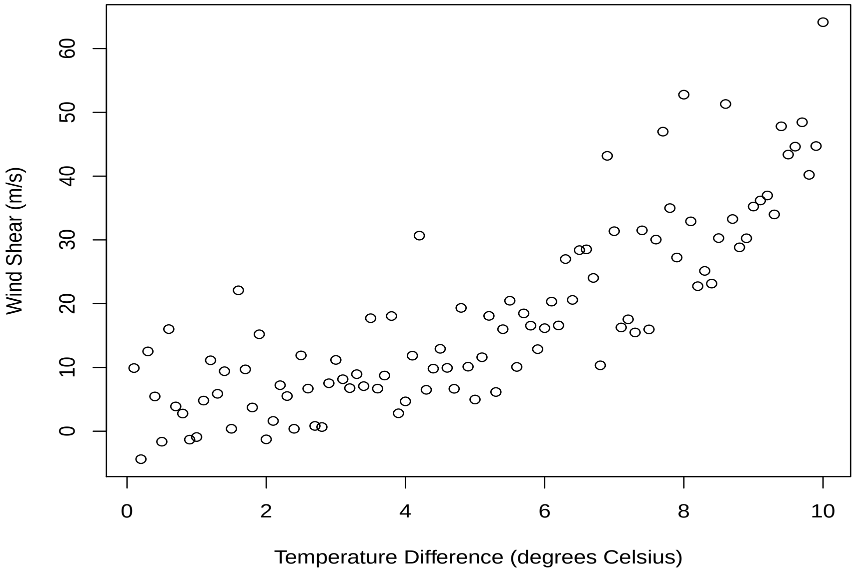


Figure 1: A scatterplot of wind shear against temperature gradient.

Our first objective is to develop a regression model to relate the distributions of the response variables Y_i to the covariates x_i using densities (1) to determine a random model component. One suggestion, based on a visual examination of curves that seem to describe the scatterplot in Figure 1, is to take the Y_i to be independent with densities $f(y_i|\xi_i, \theta)$, where f is given in (1) and, for $i = 1, \ldots, n$,

$$\xi_i = \exp(\beta_0 + \beta_1 x_i), \tag{2}$$

and where β_0 and β_1 are parameters to be estimated.

Given that extreme value distributions form location-scale families, this regression model could also be written as, for $i = 1, \ldots, n$,

$$Y_i = \exp(\beta_0 + \beta_1 x_i) + \theta\epsilon_i, \tag{3}$$

where the ϵ_i are assumed to be independent and identically distributed with densities

$$f(\epsilon) = \exp(-\epsilon)\ \exp[-\exp(-\epsilon)]; \quad -\infty < \epsilon < \infty.$$

ANSWER QUESTIONS 1 and 2 NOW

(Questions begin on page 14.)

Simultaneous maximum likelihood estimaton of β_0, β_1 and θ in the basic additive error model (3) can prove challenging in this problem. With some seemingly reasonable starting values obtained from ordinary least squares applied to the logarithm of responses and the covariates, the `R` function `nlm` returns results along with convergence code 1 which according to the `R` help file means **"relative gradient is close to zero, current iterate is probably solution"**, which is the most successful completion code available for this function. Slightly different starting values also return estimates with convergence code 1. Estimates for several starting values are given in Table 1. These values are clearly not in concert with the data of Figure 1 as the estimates of β_1 are negative. But the gradients are small for both trials, and the convergence codes indicated success for each trial.

	Estimate				Maximized
Trial	β_0	β_1	θ	Gradient	Likelihood
1	-3.11	-37.26	14.25	$(-4 \times 10^{-5}, -4 \times 10^{-6}, -6 \times 10^{-7})$	-432.461
2	-6.5	-75.9	14.64	$(-3 \times 10^{-8}, -2.9 \times 10^{-9}, -2 \times 10^{-6})$	-435.114

Table 1: Estimates and gradients returned by the function `nlm` for two different starting values (trials).

When such computational disconnects occur, one course of action is to examine likelihood "slices", in which one parameter is varied and the others held fixed. Likelihood slices in the dimensions of the three parameters β_0, β_1, and θ are presented in Figure 2.

The likelihood slices in Figure 2 clearly indicate that the dimension of β_1 is likely a cause of problems. Selecting a few values of β_1 and maximizing the log likelihood in β_0 and θ with β_1 held fixed (using a Newton-Raphson algorithm in two dimensions) verifies that the dimension of β_1 is the source of problems. If we now abandon mere slices, and use profile likelihood for β_1 we can obtain simulltaneous maximum likelihood estimates by

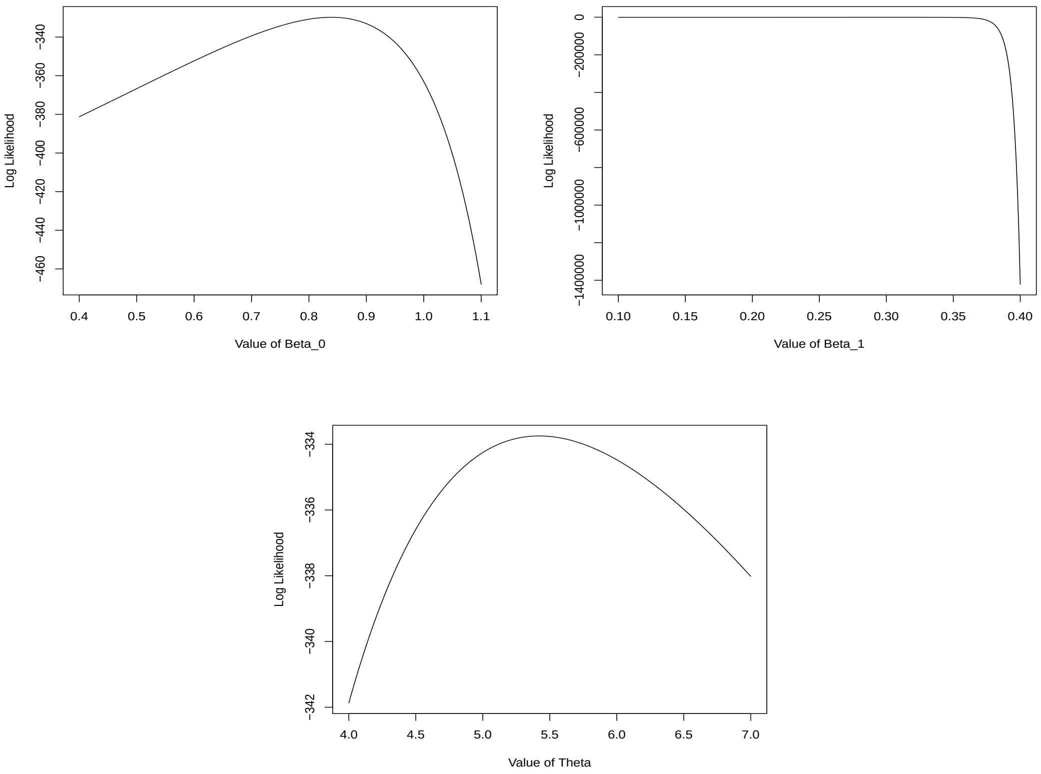


Figure 2: log likelihood slices for parameters β_0, β_1 and θ.

maximizing the profile log likelihood. This results in the estimates

$$\hat{\beta}_0 = 0.750, \qquad \hat{\beta}_1 = 0.311, \qquad \hat{\theta} = 5.107.$$

At $\beta_1 = 0.311$, the profile log likelihood is $\ell^p(\beta_1) = -320.8586$ and the inverse of the matrix of second derivatives with respect to β_0 and θ evaluated at $\hat{\beta}_0$ and $\hat{\theta}$ is

$$H^{-1} = \begin{pmatrix} 0.000750 & 0.002672 \\ 0.002672 & 0.152915 \end{pmatrix}.$$

If this matrix is used as the estimated covariance matrix for a limiting normal distribution for $\hat{\beta}_0$ and $\hat{\theta}$, approximate 95% interval estimates can be computed as

$$\begin{aligned} \beta_0 : &\quad 0.750 \pm 1.96\sqrt{0.000750} &&= (0.6964,\, 0.8037) \\ \theta : &\quad 5.431 \pm 1.96\sqrt{0.152915} &&= (4.3411,\, 5.8740) \end{aligned} \tag{4}$$

Using the general result that, for a generic scalar parameter α and its maximum likelihood estimator $\hat{\alpha}$, a profile log likelihood $\ell^p(\alpha)$ converges in distribution as

$$-2[\ell^p(\alpha) - \ell^p(\hat{\alpha})] \xrightarrow{d} \chi^2_1, \tag{5}$$

an approximate 95% interval estimate for β_1 can be computed as

$$(0.2641, 0.3382).$$

Looking again at the log likelihood slice for β_1 in Figure 2 (the right panel of the first row), we might wonder if the visual perception of that plot is affected by the scale, since the log likelihood values are so different than the values for slices in the other dimensions. If we look at this same slice, but plotting on a much more restricted range of values for β_1 we obtain the plot of Figure 3. Figure 3 would seem to indicate that although the log likelihood is flat in the dimension of β_1 globally, the log likelihood surface is actually fairly well behaved locally. The implication is that a simultaneous optimization algorithm such as Newton-Raphson should work, but is sensitive to starting values, and the starting value for β_1 needs to be quite close to the maximum likelihood estimate. Re-running the `R` function `nlm` with starting values that are quite close to the simultaneous maximum likelihood estimate in the dimension of β_1 (which we now know), we obtain the estimates,

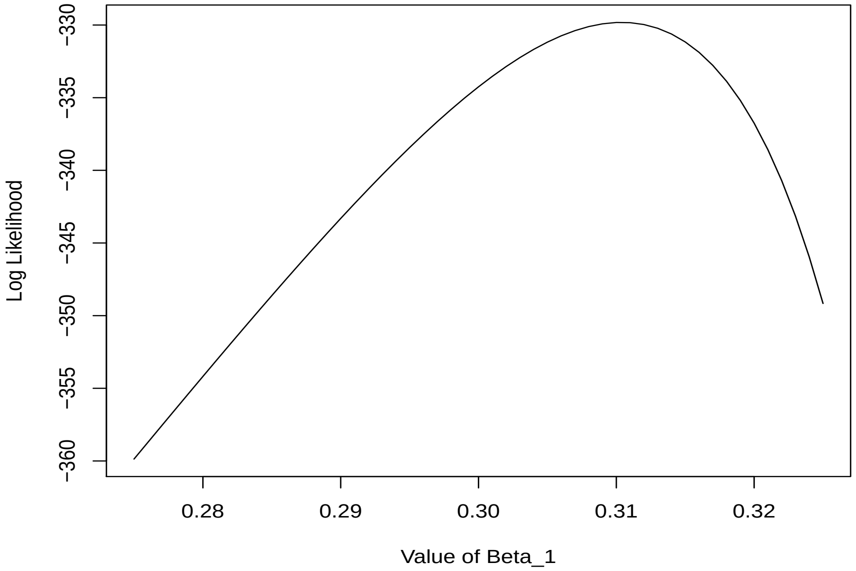


Figure 3: A slice of the log likelihood in the dimension of β_1.

$$\begin{aligned}\hat{\beta}_0 &= 0.750\\ \hat{\beta}_1 &= 0.311\\ \hat{\theta} &= 5.107\end{aligned}$$

which are identical to those obtained from the use of the profile for β_1. The gradient for these results was $(-3 \times 10^{-5},\ -3 \times 10^{-4}, 4 \times 10^{-7})$ which is actually not quite as small as those of Table 1, but the maximized log likelihood was -320.8586 which is considerably greater than the values in Table 1 and is again the same as that obtained from the profile procedure.

Using the inverse of the numerically-determined 3×3 Hessian computed by `nlm` as a covariance matrix, we obtain the following Wald theory intervals (at 95%)

$$\begin{aligned}
\beta_0 :&\quad 0.750 \pm 1.96\sqrt{0.024222} &&= (0.4450,\, 1.0551),\\
\theta :&\quad 5.108 \pm 1.96\sqrt{0.159169} &&= (4.3256,\, 5.8890),\\
\beta_1 :&\quad 0.311 \pm 1.96\sqrt{0.000324} &&= (0.2761,\, 0.3466).
\end{aligned} \tag{6}$$

Comparing the interval estimates of expressions (4) and (6) is quite striking. These intervals are reproduced in Table 2 for ease of examination. Notice from these values that the intervals for β_1 compare favorably, even though they were computed from different theoretical results (asymptotic Chi-squareness of profile log likelihood on the one hand, and asymptotic normality of maximum likelihood estimates on the other). All of the other values were computed from results of asymptotic normality. The Wald interval for θ computed in conjunction with profiling of β_1 is just a bit shorter than the interval without profiling, but the interval for β_0 computed in conjunction with profiling of β_1 is substantially shorter than the interval computed without profiling. In fact, the profile connected interval for β_0 is only 18% as wide as the interval computed from the full three-dimensional inverse information matrix.

Parameter	β_1 Profiled	β_1 Not Profiled
β_0	(0.6964, 0.8037)	(0.4450, 1.0551)
θ	(4.3411, 5.8740)	(4.3256, 5.8890)
β_1	(0.2641, 0.3382)	(0.2761, 0.3466)

Table 2: Intervals estimates computed with and without the technique of profiling for the parameter β_1.

ANSWER QUESTIONS 3 and 4 NOW

(Questions begin on page 14.)

The wind shear values in Figure 1 were measured during times at which the atmosphere was classified as "Stable." Stability in the atmosphere has to do with temperature differences at greatly different altitudes over time (as opposed to the temperature differences between not greatly different altitudes at only one point in time, which are the temperature differences in Figure 1). What is important for our consideration here is that temperature differences between 250 and 50 meters at the same time and of the magnitude of those in Figure 1 can occur with either stable or unstable atmospheric conditions. An additional data set similar to the values of Figure 1 was collected during times the atmosphere was classified as "Unstable." A scatterplot analogous to that of Figure 1 for these data is presented in Figure 4.

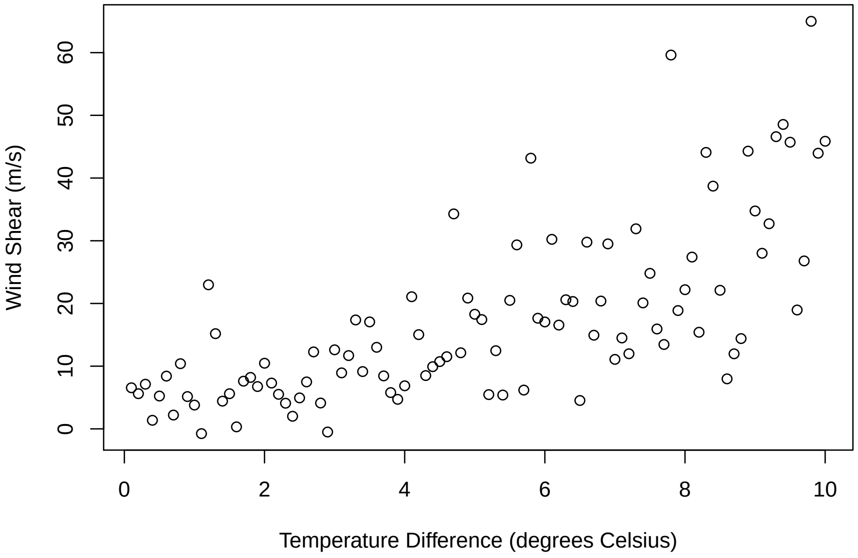


Figure 4: A scatterplot of wind shear against temperature gradient during periods of atmospheric unstability.

A question of interest to meteorologists and wind energy experts is whether the relation between wind shear and temperature differences differs in meaningful ways between times when the atmosphere is considered stable and times it is considered unstable. Model (3) was fit to the data of Figure 4 using a combination of profile (for β_1) to locate maximum likelihood estimates and then a simultaneous Newton-Raphson algorithm with values close to the maximum likelihood estimates as starting values to obtain the full 3×3 inverse observed information matrix. The resulting point and 95% Wald interval estimates are given in Table 3. The maximized log likelihood for these data collected under ustable atmospheric conditions was $\ell_u(\hat{\beta}_0, \hat{\beta}_1, \hat{\theta}) = -358.5847$.

Parameter	Point Estimate	95% Interval
β_0	1.449	(1.057, 1.841)
β_1	0.191	(0.141, 0.241)
θ	7.392	(6.234, 8.551)

Table 3: Estimates for the data of Figure 4.

To compare regressions between stable and unstable atmospheric conditions, we can conduct a likelihood ratio test of a full model in which the data sets of Figures 1 and 4 follow different response or regression functions against a reduced model in which the same response function applies to data from both figures. The maximized log likelihood for the regression with data from stable conditions (Figure 1) was $\ell_s = -320.8586$. The maximized log likelihood for the full model in our test is then

$$\ell_F = \ell_s + \ell_u = -679.4433$$

Combining the data from Figure 1 with those from Figure 4 and fitting the model yielded a maximized log likelihood for a reduced model of $\ell_R = -703.3508$ and the likelihood ratio test statistic is

$$T = -2(\ell_R - \ell_F) = 47.815$$

which results in a $p-$value of $p = 4.1 \times 10^{-11}$. Based only on this result we would reject the reduced model in favor of the full model, and conclude that stable and unstable atmospheric conditions produce different response functions in regressions of wind shear on the temperature difference at 250 and 50 meters. Scatterplots with fitted regression functions and basic residual plots of raw residuals against fitted values are shown in Figure 5 for the two data sets. Note that the left hand side scatterplots of Figure 5 are the same scatterplots as Figure 1 (top) and Figure 4 (bottom) but plotted on the same scale and with estimated response functions added. Both residual plots are also on the same scale and use the logarithm of fitted values on the horizontal axis to make visual examination easier. The visual difference in fitted response functions can be seen in Figure 5 (both scatterplots have the same range of values on the vertical axis). Right skewed distributions are also clearly depicted in these plots, which is accentuated by the fact that the response functions describe modes rather than expected values. There is also greater scatter in

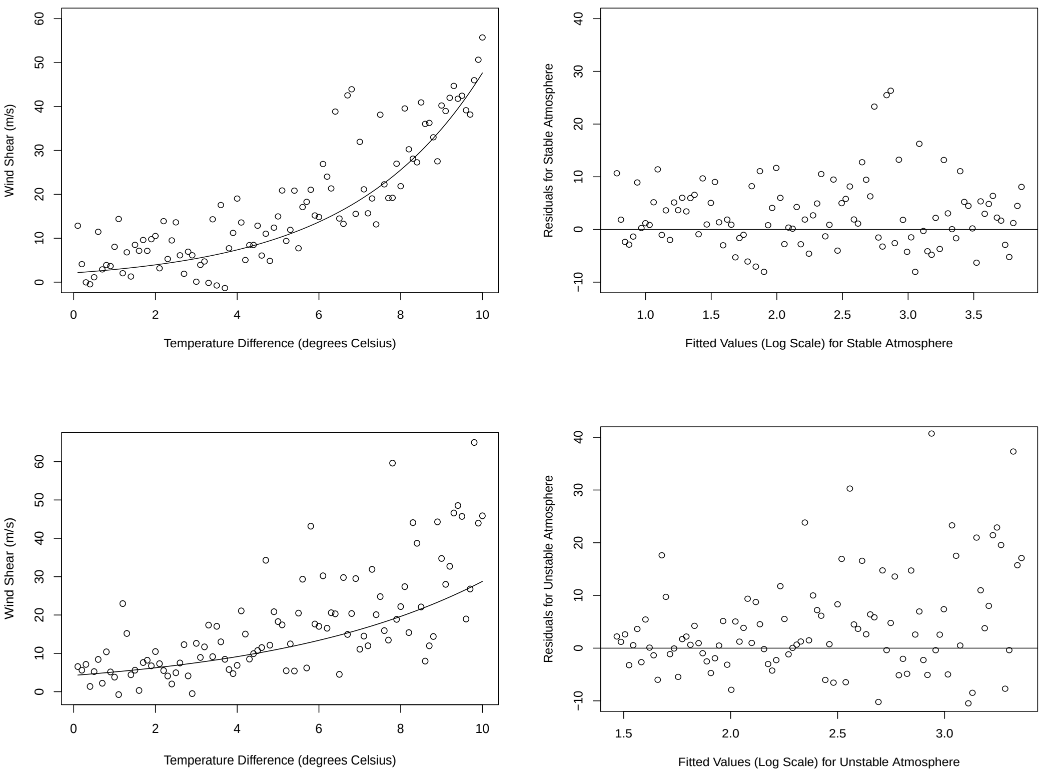


Figure 5: Fitted regressions and residual plots for data sets from stable (upper row) and unstable (lower row) atmospheric conditions.

the residuals for the unstable atmospheric situation (lower row) compared to the stable situation (upper row). It appears that there may be a bit of an increase in the spread of these residuals as fitted values get larger as well. This suggests the possibility that a model with nonconstant variances may improve the regression for data collected under unstable atmospheric conditions.

Recall that variance of the extreme value distribution (1) is proportional to the square of the parameter θ (see the text just after (1)). This, combined with the fact that our model can be written in the form of an basic additive error model (3) suggests that we might formulate a model analogous to a power of the mean model. In this case we might call our formulation a "power of the mode" model, since ξ in (1) is equal to the mode rather than the mean. This model would be, for $i = 1, \ldots, n$,

$$Y_i = \xi_i + \sigma\, \xi_i^{\phi}\, \epsilon_i, \tag{7}$$

where

$$\log(\xi_i) = \beta_0 + \beta_1 x_i,$$

and the ϵ_i are assumed to be independent and identically distributed with densities

$$f(\epsilon) = \exp(-\epsilon)\,\exp[-\exp(-\epsilon)]; \quad -\infty < \epsilon < \infty.$$

Fitting this model to data from unstable atmospheric conditions used a nested set of profiles, one for the power ϕ and one for the regression parameter β_1. Once estimates were obtained, the full 4×4 infomration matrix was computed. Point estimates and 95% Wald intervals are given for this analysis in Table 4. The scatter plot of Figure 4 is reproduced along with the estimated regression function in Figure 6. The maximized log likelihood using the estimates of Table 4 was $\ell = -313.0405$. Comparing the values of Table 4 to those of Table 3 indicates that estimates of the regression parameters β_0 and β_1 change considerably between the basic additive error model (3) and the power of the mode model (7). In fact, the intervals for both β_0 and β_1 are completely disjoint (i.e., no overlap) between the models. The fitted regession function shown in Figure 6 is even visually distinct from that in the lower left panel of Figure 5. In fact, both the point estimates and the intervals for β_0 and β_1 corresponding to model (7) fit to data from unstable conditions are quite similar to those computed under model (3) for the data collected during periods of atmospheric stability (see Table 2). The question that arises is whether data from the two situations (stable and unstable atmospheric conditions) should be considered to differ in response functions (β_0 and β_1) or be considered to have the same response function but differ in an appropriate model for variances. The first step in addressing this question would be to determine whether there is sufficient evidence to prefer model (7) to model (3) for the data from unstable atmospheric conditions.

Parameter	Point Estimate	95% Interval
β_0	0.776	(0.5695, 0.9804)
β_1	0.295	(0.2595, 0.3305)
σ	0.491	(0.2900, 0.6919)
ϕ	1.007	(0.8403, 1.1737)

Table 4: Estimates for the data of Figure 4 using model (7).

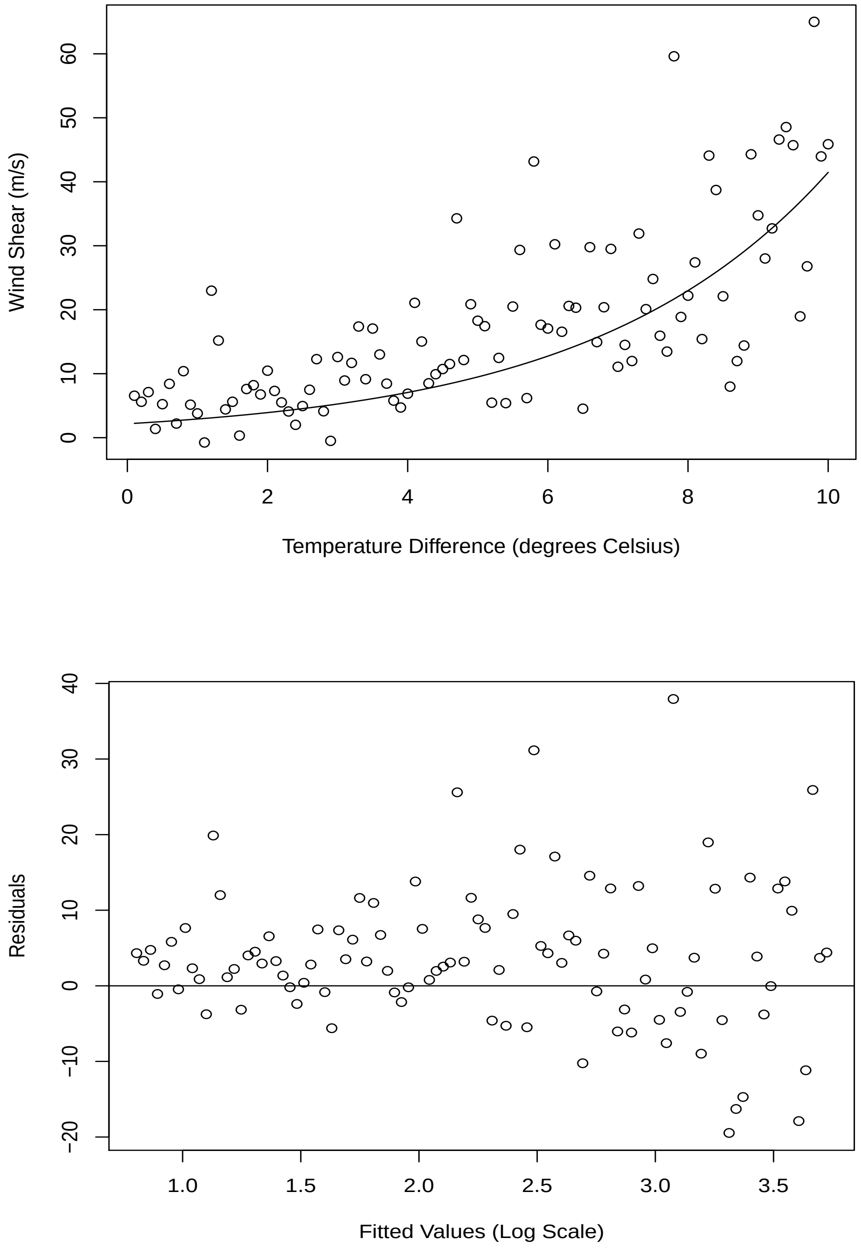


Figure 6: Power of the mode model fitted to data from periods of atmospheric unstability. Upper panel is scatterplot and fitted regression, lower panel is raw residuals.

ANSWER QUESTIONS 5, 6 and 7 NOW

(Questions begin on page 14.)

Suppose that from the analysis conducted so far we are willing to propose a final model that has a single response function for both stable and unstable atmospheric conditions. Atmospheric stability or unstability is, of course, more of a gradient of conditions than it is an actual categorical phenomenon, and is also a concept that is difficult to quantify. Consider a proposed study in which, rather than attempting to observe atmospheric stability, we desire a model in which the power ϕ in model (7) is allowed to evolve, or change slowly, over time. This study will need careful consideration of the appropriate time interval for collection of data on wind shear and temperature difference at 250 and 50 meters, but suppose that meterorologists and wind energy experts are able to determine a reasonable solution to this question. To emphasize the temporal aspect of a model for data from this proposed study, re-write the model for a sequence of observations as, for $t = 1, \ldots, T$,

$$Y_t = \xi_t + \sigma \xi_t^{\phi_t} \epsilon_t, \tag{8}$$

where

$$\log(\xi_t) = \beta_0 + \beta_1 x_t,$$

and the ϵ_t are assumed to be independent and identically distributed with densities

$$f(\epsilon) = \exp(-\epsilon)\, \exp[-\exp(-\epsilon)].$$

The concept of this model is that at times of relative atmospheric stability ϕ_t will be close to 0, while at times of relative atmospheric unstability ϕ_t will differ from 0, perhaps substantially so.

It is instructive for model (8) to examine the relevant parameter spaces explicitly. These are,

$$\begin{aligned}\beta_0 &\in (-\infty, \infty)\\ \beta_1 &\in (-\infty, \infty)\\ \sigma &\in (0, \infty)\\ \phi_t &\in (-\infty, \infty).\end{aligned}$$

Note that, while we would anticipate ϕ_t will mostly assume positive values, the parameter space includes the negative line, although values of large magnitude (either negative or positive) would lead to extremely ill-behaved distributions.

ANSWER QUESTIONS 8, 9 and 10 NOW

(Questions begin on page 14.)

This is a Summary to Help with Reference to the Numbered Models in the Question

(equation numbers are the same as given previously in the body of the question)

- Basic Additive Error Model.

  $$Y_i = \exp(\beta_0 + \beta_1 x_i) + \theta\, \epsilon_i, \tag{3}$$

  where, for $i = 1, \ldots, n$ the ϵ_i are assumed to be independent and identically distributed with densities

  $$f(e) = \exp(-\epsilon)\, \exp[-\exp(-\epsilon)]; \quad -\infty < \epsilon < \infty.$$

- Power of the Mode Model.

  $$Y_i = \xi_i + \sigma\, (\xi_i)^{\phi}\, \epsilon_i, \tag{7}$$

  where

  $$\log(\xi_i) = \beta_0 + \beta_1 x_i,$$

  and the ϵ_i are assumed to be independent and identically distributed with densities

  $$f(\epsilon) = \exp(-\epsilon)\, \exp[-\exp(-\epsilon)]; \quad -\infty < \epsilon < \infty.$$

- Dynamic Model.

  $$Y_t = \xi_t + \sigma(\xi_t)^{\phi_t} \epsilon_t, \tag{8}$$

  where

  $$\log(\xi_t) = \beta_0 + \beta_1 x_t,$$

  and the ϵ_t are assumed to be independent and identically distributed with densities

  $$f(\epsilon) = \exp(-\epsilon)\, \exp[-\exp(-\epsilon)].$$

Questions

1. The form of the regression model as a basic additive error model as given in (3) would seem to suggest that one could use generalized least squares to estimate the parameters β_0 and β_1. That is, consider minimizing in β_0 and β_1 the quantity

$$Q = \sum_{i=1}^{n} w_i \left[Y_i - exp(\beta_0 + \beta_1 x_i)\right]^2,$$

   for appropriately chosen weights w_i.

   Explain why this might not be such a good idea.

   Hint: the answer has nothing to do with the weights w_i in Q given above.

2. Write the log likelihood and its first derivatives with respect to β_0, β_1 and θ for the basic additive error model (2).

   Hint: It might be convenient to let $z_i = (1/\theta)(y_i - \xi_i)$

3. Using expression (5) define the set of β_1 values that are used to produce the interval (0.2641, 0.3382).

4. Both of the intervals for β_0 in Table 2 on page 6 of the question are computed as Wald intervals, that is, using the square root of the $[1, 1]$ diagonal element of an inverse Hessian matrix as the standard error of the estimate. Given that the maximum likelihood estimates are the same, explain why the standard errors and thus the intervals computed from them are so dramatically different.

5. Setting $\phi = 0$ in the power of the mode model (7) results in the basic additive error model (3). A likelihood ratio test in which model (3) is the reduced model and model (7) is the full model would result in

$$-2[(-358.5847) - (-313.0405)] = 91.0884,$$

   which has an associated $p-$value so small it can be considered to be 0. Thus, we would prefer the power of the mode model to the basic additive error model for describing the data collected under unstable atmospheric conditions. Discuss the degree to which this then indicates that the power of the model model should be considered as a good representation of whatever mechanism generated the actual data in this problem.

   Hint: Consider the motivation for development of the power of the mode model in the first place.

6. Raw residuals from the fit of the power of the mode model (7) to data from unstable conditions in Figure 6 are perhaps somewhat more visually pleasing than those from the fit of model (3) in the lower right panel of Figure 5. But they have little use in determining whether or not the power of the mode model does a good job of accounting for unequal variances in the data. Suggest how one might construct another type of residuals that are more useful than raw residuals for this purpose, and how one might use them in an assessment of the models.

   Hint: consider the overall question of the appropriateness of models (7) and (3) for data collected during unstable atmospheric conditions.

7. Suppose we are willing to accept that the basic additive error model (3) with constant θ is adequate to describe the data collected during conditions of atmospheric stability, and that the power of the mode model (7) is adequate to describe the data collected during conditions of atmospheric unstability. We would like a formal procedure to assess whether the response functions of these two models, namely,

$$\log(\xi_i) = \beta_0 + \beta_1 x_i,$$

   can be considered the same for the two data sets, or whether they should not be considered the same for the two data sets. If a likelihood ratio test could be used as such a formal procedure, we need full and reduced models that are nested. An obvious full model here would consist of model (3) for the data from stable conditions and model (7) for data from unstable conditions, and assuming independence between all response variables invovled. Explicitly write a possible reduced model for this procedure, or indicate that no appropriate reduced model exists (and why).

8. Propose a possible model for values of ϕ_t in the dynamic model (8). Keep in mind that observations in the study under consideration are likely to be fairly close in time (e.g., maybe hourly).

   Hint: We are not able to verify any possible model at this point, only propose. Any model used in the potential study would need careful assessment if used.

9. Suggest a prior distribution that might be used in a Bayesian analysis of the model you completed in Question 8. You will need to indicate how you will determine a joint prior, and also specify any components of that joint prior (i.e., specify particular densities).

10. Suppose that we will conduct an analysis by simulating from the joint posterior distribution using an overall Gibbs Sampling algorithm. First give a form for the joint posterior (up to a proportionality constant) that we wish to simulate from, and then give the form of full conditional posterior distributions (up to a proportionality constant) that will be needed for this MCMC algorithm. Use the standard notation that $p(q|\cdot)$ denotes the full conditional posterior distribution of a quantity q. You may use $p(\cdot)$ and $\pi(\cdot)$ as generic notation for probability density functions of whatever the argument is, so that $p(q_1|\cdot)$ and $p(q_2|\cdot)$ denote the full conditional posterior distributions of quantities q_1 and q_2, but may not be of the same form.

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. Least squares methods, including generalized least squares, are defined as the solution to minimization problems in vector spaces. They have statisitcal properties that are related to the use of squared error loss as a criterion by which to measure estimation error which, in regression problems, involves expectation functions. That is, generalized least squares assumes the expectation function is given by the location portion of the location-scale transformation of additive errors that defines the model. Here, that location transformaton does not give an expectation function but, rather, a function for modes of response distributions. Estimating the regression parameters β_0 and β_1 assuming they describe an expectation function will result in over-estimates (given that the response distributions are right skew). If those estimates are assumed to describe modes as in model (3) estimation of the response distributions at given covariate values will be of poor quality. If, as indicated in the description of the problem, estimation of the probability of exceeding a given response value (35 m/s in the question) is important, this would be adversely affected. *Note: Generalized least squares estimates for the data of Figure 1 in the question produce a fitted regression curve lying entirely above that shown in the upper left panel of Figure 5 of the question.*

Question 2. Following the hint, let $z_i = (1/\theta)(y_i - \xi_i)$. The marginal density of response variable Y_i is then

$$f(y_i|\xi_i, \theta) = \frac{1}{\theta}\,\exp(-z_i)\,\exp[-\exp(-z_i)],$$

and by independence the log likelihood is then

$$\ell(\beta_0, \beta_1, \theta) = \sum_{i=1}^{n} \ell_i = -\sum_{i=1}^{n}(z_i) - \sum_{i=1}^{n}[\exp(-z_i)] - n\log(\theta).$$

Also, let $\eta_i = \beta_0 + \beta_1 x_i$. First derivatives are most easily determined through use of the chain rule as

$$\begin{aligned}
\frac{\partial \ell_i}{\partial \beta_0} &= \frac{\partial \ell_i}{\partial z_i}\frac{\partial z_i}{\partial \xi_i}\frac{\partial \xi_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_0} \\
\frac{\partial \ell_i}{\partial \beta_1} &= \frac{\partial \ell_i}{\partial z_i}\frac{\partial z_i}{\partial \xi_i}\frac{\partial \xi_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_1} \\
\frac{\partial \ell_i}{\partial \theta} &= \frac{\partial \ell_i}{\partial z_i}\frac{\partial z_i}{\partial \theta} - \frac{1}{\theta},
\end{aligned}$$

where the last term in the partial for θ comes from the first factor of $1/\theta$ in the densities. The particular terms in these derivatives are

$$\begin{aligned}
\frac{\partial \ell_i}{\partial z_i} &= \exp(-z_i) - 1 \\
\frac{\partial z_i}{\partial \xi_i} &= -\frac{1}{\theta} \\
\frac{\partial \xi_i}{\partial \eta_i} &= \exp(\eta_i) \\
\frac{\partial \eta_i}{\partial \beta_0} &= 1 \\
\frac{\partial \eta_i}{\partial \beta_1} &= x_i \\
\frac{\partial z_i}{\partial \theta} &= -\frac{(y_i - \xi_i)}{\theta^2}
\end{aligned}$$

Question 3. The values of β_1 contain in the interval are defined as

$$\{\beta_1 : -2[\ell^p(\beta_1) - \ell^p(\hat{\beta}_1)] \geq \chi^2_{1,0.95}\},$$

where $\chi^2_{1,0.95}$ is the 0.95 quantile of a Chi-square distribution with 1 degree of freedom.

Question 4. The interval connected with the profile likelihood procedure is computed from a 2×2 inverse information matrix that excludes derivatives with respect to β_1. The other interval is computed from a 3×3 inverse information matrix that includes derivatives with respect to β_1. Thus, the interval connected with the profile procedure ignores the effect of uncertainty in $\hat{\beta}_1$ on uncertainty in $\hat{\beta}_0$ (and $\hat{\theta}$ for that matter). If the estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are highly correlated, which we might well expect since β_0 is an intercept parameter and β_1 is essentially a slope parameter (within the context of a log link), this could cause the observed difference in the widths of the two intervals. We might also surmise that $\hat{\beta}_1$ is much less correlated with $\hat{\theta}$ because the two intervals for θ are not dramatically different.

Note: this is, in fact, true. Computed from the full 3×3 *inverse information matrix, the correlation between* $\hat{\beta}_0$ *and* $\hat{\beta}_1$ *is* -0.98*, while that between* $\hat{\beta}_1$ *and* $\hat{\theta}$ *is* -0.20*.*

Question 5. The likelihood ratio test poses a choice between the basic additive error model (3) and the power of the mode model (7) for these data. It does not necessarily demonstrate that either model is appropriate for the data. The motivation for development of the power of the mode model was an indication of unequal spread in a plot of raw residuals (Figure 5). Nothing in the likelihood ratio test or the residual plot of Figure 6 address the question of whether the power of the mode model was effective in dealing with this apparent unequal variance.

Question 6. Generalized residuals would be valuable to address the overall adequacy of the models for fitting the observed data. Generalized residuals would be defined here as,

$$u_i = \int_{-\infty}^{y_i} f(t|\hat{\beta}_0, \hat{\beta}_1, \hat{\theta})\ dt$$

where $f(\cdot)$ is the extreme value density of expression (1) in the exam with the parameter ξ replaced by $\hat{\xi}_i$ and the parameter θ replaced by $\hat{\theta}$. Here,

$$\hat{\xi}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

To use the generalized residuals $\{u_i : i = 1, \ldots, n\}$ as a diagnostic we could plot their empirical distribution functions under the two models and compare each to a theoretical distributon function of the uniform distribution on (0, 1). To use the residuals in a more formal procedure, we could conduct a test of uniformity for each model.

Note: one could attempt to standardize or studentize the raw residuals for these models, but that is complicated by the fact that we do not actually have estimates of the expected values.

Question 7. A formal procedure to address this question could take the form of a likelihood ratio test in which the full model consists of fitting model (3) to data from stable conditions and model (7) to data from unstable conditions (and then adding the maximized log likelihoods). An appropriate reduced model is the same thing with the constraint that β_0 and β_1 are the same in the two groups of data. One way an appropriate reduced model can be formulated would be by combining the data (and indexing consecutively $i = 1, \ldots, n$) after defining an indicator, v_i say, where

$$v_i = \begin{cases} 1 & \text{if } Y_i \text{ is from unstable conditions} \\ 0 & \text{otherwise} \end{cases}$$

The reduced model could then be written as

$$Y_i = \xi_i + [\theta(1 - v_i) + \sigma\, v_i](\xi_i)^{v_i\,\phi}\epsilon_i,$$

where $\log(\xi_i) = \beta_0 + \beta_1 x_i$ and the ϵ_i are taken to be independent and identically distribued with densities

$$f(\epsilon_i) = \exp(-\epsilon_i)\, \exp[-\exp(-\epsilon_i)].$$

The test would have 2 degrees of freedom (corresponding to different or the same regression functions for the two types of data). Note that this test could be considered to fall into the category of "reversed hypothesis" testing. Our scientific hypothesis is that there is really only one regression function, and yet that structure is embodied in the reduced model (the null hypothesis).

Question 8. One possible model for the ϕ_t would be to assign these data model parameters a first-order autoregressive structure,

$$\phi_t = \alpha\, \phi_{t-1} + w_t,$$

where the w_t are taken as independent and identically distributed random variables having $N(0,\, \tau^2)$ distributions. Note that the joint distribution of the ϵ_t for $t = 1, \ldots, T$ can be derived explicitly based on conditioning, and the assumption that ϵ_0 has a normal distribution with mean 0 and the variance of the marginal stationary distribution.

Question 9. Let $\pi(\cdot)$ denote a generic prior distribution (so $\pi(x)$ is the density of X and $\pi(y)$ is the density of Y, although these are not the same formulas). We need a joint prior for β_0, β_1, σ, α, and τ^2, which we could formulate as a product form,

$$\pi(\beta_0, \beta_1, \sigma, \alpha, \tau^2) = \pi(\beta_0)\,\pi(\beta_1),\; \pi(\sigma)\,\pi(\alpha)\,\pi(\tau^2).$$

Alternatively, we could specify the prior for β_0 and β_1 jointly, since we have information that indicates the estimates of these parameters will contain a high degree of correlation.

The individual components might be chosen as

$$\begin{array}{rcl}
\pi(\beta_0) & \text{is} & N(0,\, \lambda_0) \\
\pi(\beta_1) & \text{is} & N(0,\, \lambda_1) \\
\pi(\sigma) & \text{is} & Unif(0,\, A) \\
\pi(\tau^2) & \text{is} & Unif(0,\, B) \\
\pi(\alpha) & \text{is} & N(0,\, \lambda_\alpha) \text{ truncated below at } -1 \text{ and above at } 1 \\
 & \textit{or} & \\
\pi(\alpha) & \text{is} & Unif(-1,\, 1)
\end{array}$$

Here, λ_0, λ_1, and λ_α would be chosen fairly large to give diffuse priors. The upper limits on the uniform priors are easily changed in a sensitivity analysis. The range of α is restricted to make the autoregressive process for ϕ_t stationary.
Other choices are possible, as long the support results in the data model parameter spaces given in the question.

Question 9. We wish to make inference based on the joint posterior $p(\beta_0, \beta_1, \sigma, \alpha, \tau^2 |\, \boldsymbol{y})$,, which can be accomplished by simulating from

$$p(\beta_0, \beta_1, \sigma, \alpha, \tau^2, \boldsymbol{\phi} |\, \boldsymbol{y}) = p(\beta_0, \beta_1, \sigma, \alpha, \tau^2, \phi_0, \phi_1, \ldots, \phi_T |\, \boldsymbol{y}).$$

The full conditional posteriors needed for this algorithm would be

$$p(\beta_0|\cdot) \;\propto\; \pi(\beta_0)\, f(\boldsymbol{y}|\beta_0, \beta_1, \sigma, \boldsymbol{\phi})$$

$$
\begin{aligned}
p(\beta_1|\cdot) &\propto \pi(\beta_1)\, f(\boldsymbol{y}|\beta_0, \beta_1, \sigma, \boldsymbol{\phi}) \\
p(\sigma|\cdot) &\propto \pi(\sigma)\, f(\boldsymbol{y}|\beta_0, \beta_1, \sigma, \boldsymbol{\phi}) \\
p(\alpha|\cdot) &\propto \pi(\alpha)\, g(\boldsymbol{\phi}|\alpha, \tau^2) \\
p(\tau^2|\cdot) &\propto \pi(\tau^2)\, g(\boldsymbol{\phi}|\alpha, \tau^2) \\
p(\phi_0|\cdot) &\propto g(\phi_0|\alpha, \tau^2)\, g(\phi_1|\alpha, \tau^2, \phi_0)
\end{aligned}
$$

and, for $t = 1, \ldots, T$

$$p(\phi_t|\cdot) \propto g(\phi_t|\alpha, \tau^2, \phi_{t-1})\, g(\phi_{t+1}|\alpha, \tau^2, \phi_t)\, f(y_t|\beta_0, \beta_1, \sigma, \phi_t).$$

Note that there are two rather subtle pieces to this.

(a) We are assuming that the process has existed for quite awhile by the time we take our first observation corresponding to Y_1. Thus, the use of the conditional distribution $g(\phi_1|\phi_0)$ in formulating

$$g(\boldsymbol{\phi}|\alpha, \tau^2) = g(\phi_0|\alpha, \tau^2) \prod_{t=1}^{T} g(\phi_t|\alpha, \tau^2, \phi_{t-1}),$$

and we assume that ϕ_0 can be considered to have the common marginal stationary distribution of the ϕ_t which will be $N(0, \tau^2/(1-\alpha^2))$.

(b) All of the full conditional posteriors are proportional to

$$\pi(\beta_0, \beta_1, \sigma, \alpha, \tau^2)\, f(\boldsymbol{y}|\beta_0, \beta_1, \sigma, \boldsymbol{\phi})\, g(\boldsymbol{\phi}|\alpha, \tau^2).$$

Using the product form of the joint prior, conditional independence of the Y_t, and the first-order Markov property assumed for the ϕ_t this becomes

$$\pi(\beta_0)\, \pi(\beta_1)\, \pi(\sigma)\, \pi(\alpha)\, \pi(\tau^2) \left[\prod_{t=1}^{T} f(y_t|\beta_0, \beta_1, \sigma, \phi_t)\right] \left[g(\phi_0|\alpha, \tau^2) \prod_{t=1}^{T} g(\phi_t|\alpha, \tau^2, \phi_{t-1})\right],$$

from which the forms for $p(\phi_0|\cdot)$ and $p(\phi_t|\cdot)$ can be determined directly.