An animal in the wild can be located when it has a small radio transmitter attached to it. An antenna picks up the signal from the transmitter. Signal strength provides information about the angle and distance from the antenna to the animal. Eventually, this information will be used to locate the animal.

Data were collected by measuring signal strength when the transmitter was at a known location in the field. Figure 1 shows the setup. `a` is the angle between the antenna direction and the transmitter, in radians. Positive angles are clockwise from the antenna direction; negative angles are counterclockwise. `d` is the distance between the antenna and transmitter, in meters. Signal strength is known to be largest when the transmitter is directly in front of the antenna (angle = 0) and to decline with increasing deviation, either clockwise or counterclockwise, from the antenna direction. Signal strength is also known to decrease with increasing distance from the antenna, when measured at the same angle.

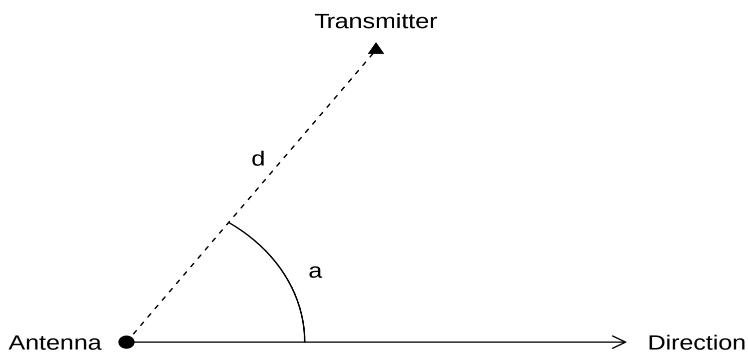

Figure 1: Diagram showing the relationship between antenna pointing direction and transmitter location.

Data were collected by randomly choosing an angle and measuring signal strength when the transmitter was 25, 50, 100, 150, and 200 meters from the antenna at the selected angle. Two signal strength measurements were obtained at each of the five distances for each selected angle. The measurement process was carried out for a total of 10 different angles.

Let Y_{ijk} be signal strength in decibels (dB) for the ith angle, jth distance, and kth measurement $(i = 1, \ldots, 10, j = 1, \ldots, 5, k = 1, 2)$. The correspondence of i and j indices to specific angles and distances is given in Table 1 on page 2, which gives the number of observations available for analysis for each combination of angle and distance. Because the measurement system does not report signal strength when the signal is too weak, some counts in Table 1 are zero. Until stated otherwise, these missing values will be ignored in the analysis and treated as if measurement for that angle and distance was never attempted. There are a total of 84 non-missing observations in the first data set. Rows of Table 1 are sorted in order of increasing angle. The index i indicates the order of measurement, i.e., i=1 was the first angle to be measured, i=2 was the second, up to i = 10 was the last.

Table 1: Counts of signal strength observations by angle and distance.

Angle, in radians	Distance, in meters 25 ($j=1$)	50 ($j=2$)	100 ($j=3$)	150 ($j=4$)	200 ($j=5$)
-1.48 ($i=7$)	2	2	2	0	0
-0.92 ($i=4$)	2	2	2	2	0
-0.62 ($i=5$)	2	2	2	2	2
-0.47 ($i=2$)	2	2	2	2	2
-0.13 ($i=1$)	2	2	2	2	2
-0.12 ($i=10$)	2	2	2	2	2
0.76 ($i=9$)	2	2	2	2	2
0.90 ($i=6$)	2	2	2	2	0
1.31 ($i=8$)	2	2	2	0	0
1.44 ($i=3$)	2	2	2	0	0

Part I

The first analyses of these data are based on the additive main effects model:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \gamma_j + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2), \end{aligned} \tag{1}$$

where $\mu, \alpha_1, \ldots, \alpha_{10}, \gamma_1, \ldots, \gamma_5 \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown parameters.

1. Describe, with reference to this study, what $\mu + \frac{\Sigma_{i=1}^{10}\alpha_i}{10} + \gamma_5$ represents.

2. Is $\mu + \frac{\Sigma_{i=1}^{10}\alpha_i}{10} + \gamma_5$ estimable? Show why or why not.

3. The error sums-of-squares (SSEs) for model (1) and each of 3 submodels are given in Table 2 on page 3. All models assume error terms are independent and normally distributed with mean zero and constant variance. Consider the Type III F test of $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = \gamma_4 = \gamma_5$.

a) Compute the value of the test statistic for this test.
b) State the distribution of the test statistic under the null hypothesis.
c) Using the provided distribution tables, report an approximate p-value.
d) Write a one-sentence conclusion for this test in the context of this problem.

If additional information is needed for any part of this problem, state what is needed.

4. Each SSE in problem **3** can be written as $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P_X})\boldsymbol{Y}$, where $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'$ for an appropriate matrix $\boldsymbol{X}$. Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{X}_3$, and $\boldsymbol{X}_4$ be such matrices for models (1), (2), (3), and (4), respectively. Each has an associated projection matrix: $\boldsymbol{P}_{\boldsymbol{X}_1}$, $\boldsymbol{P}_{\boldsymbol{X}_2}$, $\boldsymbol{P}_{\boldsymbol{X}_3}$, and $\boldsymbol{P}_{\boldsymbol{X}_4}$. Is the sum of squares $\boldsymbol{Y}'(\boldsymbol{P}_{\boldsymbol{X}_1} - \boldsymbol{P}_{\boldsymbol{X}_4})\boldsymbol{Y}$ independent of the SSE for model (1) given by $\boldsymbol{Y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}_1})\boldsymbol{Y}$? Explain.

Table 2: Error sums-of-squares for the additive model (1) and three submodels.

Model number	$\mathrm{E}\, Y_{ijk}$	SSE	
1	$\mu + \alpha_i + \gamma_j$	42	
2	$\mu + \alpha_i$	6515	(2)
3	$\mu + \gamma_j$	6780	(3)
4	μ	19868	(4)

For cells in Table 1 with positive counts, a cell-means model for the combinations of angle and distance with data is:

$$
\begin{aligned}
Y_{ijk} &= \mu_{ij} + \varepsilon_{ijk}, \\
\varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2).
\end{aligned} \tag{5}
$$

SSE for model (5) is 27.

5. Consider a test of whether model (5) fits the data significantly better than model (1). Describe, in terms of angle and distance effects on signal strength, the null hypothesis that is being tested by the comparison of these two models.

6. Carry out the test described in problem **5**.

a) Compute the value of your test statistic.
b) State its distribution under the null hypothesis.
c) Using the provided distribution tables, report an approximate p-value.
d) Write a one-sentence conclusion in terms of angle and distance effects.

If additional information is needed for any part of this problem, state what is needed.

Part II

Antenna physics suggests that the signal strength received at the antenna will decline linearly with the distance to the transmitter. This suggests the model

$$
\begin{aligned}
Y_{ijk} &= \beta_0 + \alpha_i + \beta_1 d_j + \varepsilon_{ijk}, \\
\varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2),
\end{aligned} \tag{6}
$$

where d_j is the distance for observation Y_{ijk} and β_1 is the slope relating the change in average signal strength to the change in distance.

This model is fit to the data from **Part I** by ordinary least squares (OLS). SSE for model (6) is 45.

7. The ANOVA table corresponding to model (6) is

Source	df	Type III SS
Distance	?	?
Angle	?	13353
Error	?	45

a) What are the degrees of freedom for distance, for angle, and for error?
b) What is the type III SS for distance?

Note: You may need information provided earlier about models (1), (2), (3), and (4).

8. Model (6) assumes that the effect of distance is linear. Using information from **Parts I** and **II**, construct a test to evaluate this assumption.

a) Report the value of your test statistic.
b) State its distribution under the null hypothesis.
c) Using the provided distribution tables, report an approximate p-value for this hypothesis test.
d) Write a one-sentence conclusion for this test.

If additional information is needed for any part of this problem, state what is needed.

9. For this data set, the SSE for model (6) is larger than the SSE for model (1). Will this always happen, i.e., is this a general result that applies to any data set with the combinations of angle and distance shown in Table 1? Carefully explain why or why not.

10. Figures 2 and 3 on page 5 show model diagnostics for model (6). What do these plots suggest about the appropriateness of model (6)?

Mean signal strengths are estimated under Model (6) at various combinations of angle and distance. Estimated values and their standard errors for some combinations of angle and distance are shown in Table 3.

Table 3: Estimated mean signal strength and its standard error for three combinations of angle and distance.

Angle	Distance	$\widehat{Y}_{ijk}$	se $\widehat{Y}_{ijk}$
0.47	25	81.77	0.34
-0.92	50	68.54	0.36
-0.13	100	70.11	0.38

11. Construct a 95% confidence interval for the mean signal strength at an angle of -0.92 and a distance of 50.

12. There is a new measurement of signal strength at an angle of -0.92 and a distance of 50. The observed signal strength is 70.0. Because this is far outside the confidence interval in problem **11**, the researcher suspects something has gone wrong with the equipment. Do you agree? Briefly explain why or why not.

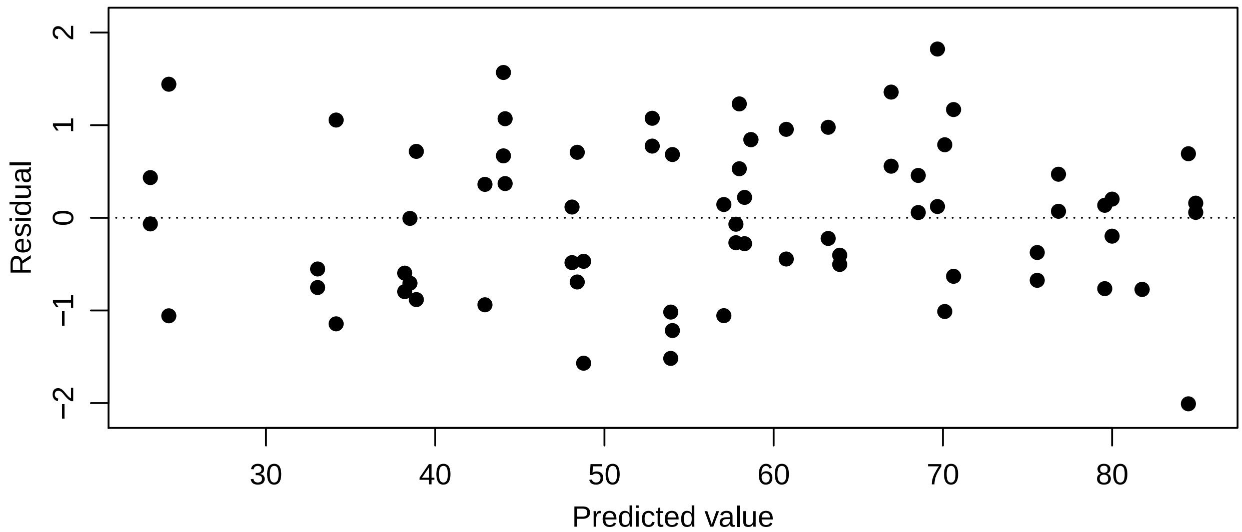

Figure 2: Residuals and predicted values from fitting model (6).

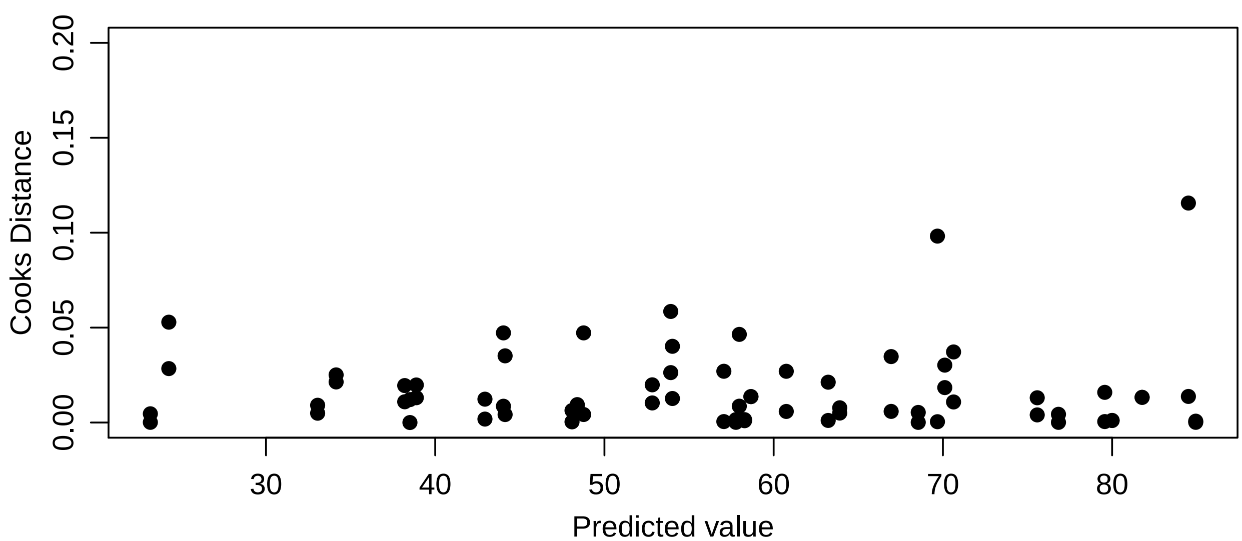

Figure 3: Cook's Distance and predicted values from fitting model (6).

Part III

Antenna physics suggests that the signal strength received at the antenna, at any fixed distance d_j, can be modeled parametrically as

$$\begin{aligned} Y_{ijk} &= \beta_{0j} + \beta_2 \cos(a_i) + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2), \end{aligned}$$

where the angle between the antenna direction and the transmitter is a_i and β_{0j} is a function of d_j.

The purpose behind modeling dependence of signal strength on angle and distance is to estimate the transmitter location from a collection of signal strength measurements. Data are to be collected from 4 towers, each with 4 antennae. Figure 4 shows a picture of the full physical setup, involving 4 antennae at each of 4 corners of a rectangular region.

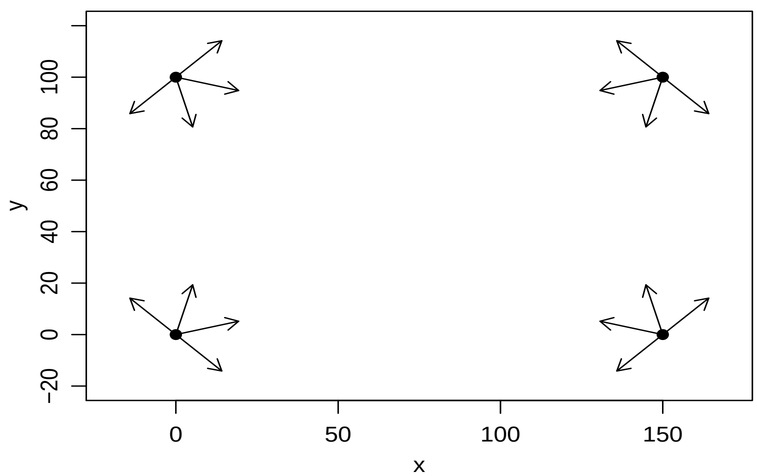

Figure 4: Full setup used to collect data to estimate an unknown location (x, y). Four towers are at the corners of the study area and are indicated by dots. Each tower has four antennae. The direction for each antenna is indicated by an arrow.

Table 4 shows a typical data set. Note that signal strength measurements are missing for some antennae. For all problems in this part, suppose values are missing because the corresponding antennae were not turned on.

Table 4: Typical data set showing signal strength for each of the 16 antennae. Signal strength was not reported from some antennae, as indicated by a dash (–) in the table.

	Antenna			
Tower	1	2	3	4
1	53	69	55	–
2	–	70	77	56
3	–	59	64	–
4	–	66	71	–

Now that we have signal strength measurements from 4 antennae on each of 4 towers for a transmitter at a single location, we need completely new notation. Define Y_{ij} as the signal strength measured at antenna j on tower i, the function $a_{ij}(x, y)$ as the angle between the unknown location (x, y) and the direction of antenna j on tower i, and the function $d_i(x, y)$ as the distance between the unknown location (x, y) and tower i. The effect of angle and the effect of distance will be assumed to be additive.

13. Based on what you have been told and have learned from your analyses in **Parts I**, **II**, and **III**, write down an appropriate model for signal strength, Y_{ij}, at tower i and antenna j as a function of the unknown location (x, y). Explain your choice.

14. Describe, in no more than a few sentences, an appropriate statistical method to estimate (x, y) based on your model in problem **13**.

15. For a sufficiently large number of towers, the estimated location $\boldsymbol{l} = (\hat{x}, \hat{y})'$ is approximately bivariate normal with mean at the true location, $\boldsymbol{\mu} = (\mu_x, \mu_y)'$, and covariance matrix estimated by a matrix $\widehat{\Sigma}$. The data in Table 4 were collected from a transmitter at an unknown location. The biologists using the system want to know whether the data are consistent with the transmitter being at a specified location, $\boldsymbol{\kappa} = (x^*, y^*)'$. Write down, but do not evaluate, an appropriate test statistic for testing the null hypothesis that the transmitter is at that specified location, i.e., $\boldsymbol{\mu} = \boldsymbol{\kappa}$.

Part IV

The analyses in **Parts I**, **II**, and **III** have ignored cases where measurements were actually attempted but not reported because the signal strength was too low. The reporting limit is known; signal strength less than 50 dB is not reported. Table 5 shows the data including censored values.

Table 5: Typical data set showing signal strength for the 16 antennae with 6 censored values.

	Antenna			
Tower	1	2	3	4
1	53	69	55	< 50
2	< 50	70	77	56
3	< 50	59	64	< 50
4	< 50	66	71	< 50

16. Maximum likelihood can be used to estimate the location when the data includes censored values. Assume that the observed signal strength, Y_{ij} for tower i and antenna j, can be written as $Y_{ij} = f(x, y) + \varepsilon_{ij},\ \ \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$, where $f(x, y)$ is your model from problem **13**. Write down, but do not numerically evaluate, the contribution to the log likelihood for

a) The observation for tower 1, antenna 1, with measured signal strength of 53.

b) The observation for tower 1, antenna 4, with a censored value of < 50.

1. $\mu + \frac{\Sigma_{i=1}^{10}\alpha_i}{10} + \gamma_5$ is the average signal strength at distance 200m, averaged over the angles used in the study.

2. Yes, $\mu + \frac{\Sigma_{i=1}^{10}\alpha_i}{10} + \gamma_5$ is estimable.
A linear combination of parameters, $\boldsymbol{C\beta}$ is estimable if it can be expressed as a linear combination of expected values of observation, $\boldsymbol{A}\mathrm{E}\,Y$, for some vector $\boldsymbol{A}$. There are 5 angles with observed signal strength at 200m. For these angles, E $\boldsymbol{Y} = \mu + \alpha_i + \gamma_5$. Under the additive model, E $Y_{i3k} - Y_{i5k} = \gamma_3 - \gamma_5$ for any i, so for the 5 angles without observed signal strength, E $Y_{i5k} =$ E $Y_{i3k} +$ E $Y_{25k} -$ E Y_{23k}. Hence, the desired quantity is a linear combination of expected values of observed cells, so it estimable.
Notes: The argument is equally valid using rows of $j = 1$ (distance of 25m), $j = 2$ (distance of 50m), or $j = 3$ (distance of 100m).
$\mu + \frac{\Sigma_{i=1}^{10}\alpha_i}{10} + \gamma_5$ is not estimable under the more general cell means model, because then E $Y_{i3k} - Y_{i5k}$ is not a constant.

3. Note: The type III F test is the comparison of models (1) and (2).

- **a**) F = 2697, Note: computed as $\frac{(6515-42)/(74-70)}{42/70}$
- **b**) Central F distribution with 4, 70 degrees of freedom
- **c**) $p < 0.0001$.
- **d**) Very strong evidence of at least one difference between the distance means.

4. Yes, these two sum-of-squares are independent under model (1). Each is a quadratic form. Two quadratic forms, $\boldsymbol{Y}'\boldsymbol{AY}$ and $\boldsymbol{Y}'\boldsymbol{BY}$ are independent if $\boldsymbol{AB} = 0$. $\boldsymbol{A} = (\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_4}})$ and $\boldsymbol{B} = (\boldsymbol{I} - \boldsymbol{P_{X_1}})$. $\boldsymbol{AB} = (\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_4}})(\boldsymbol{I} - \boldsymbol{P_{X_1}})$ = $\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_4}} - \boldsymbol{P_{X_1}P_{X_1}} + \boldsymbol{P_{X_1}P_{X_4}}$. Each $\boldsymbol{P_{X_i}}$ is idempotent, so $\boldsymbol{P_{X_1}P_{X_1}} = \boldsymbol{P_{X_1}}$. The column space of $\boldsymbol{X}_4$ is a subspace of $\boldsymbol{X}_1$, so $\boldsymbol{P_{X_1}P_{X_4}} = \boldsymbol{P_{X_4}}$. Hence $\boldsymbol{AB} = \boldsymbol{P_{X_1}} - \boldsymbol{P_{X_4}} - \boldsymbol{P_{X_1}} + \boldsymbol{P_{X_4}} = 0$.

5. The null hypothesis is that there is no interaction between angle effects and distance effects. Note: the column space of model (5) is the same as the column space of the model with an angle by distance interaction, i.e. E $Y_{ijk} = \mu + \alpha_i + \gamma_j + \alpha\gamma_{ij}$.

6.
- **a**) F = 0.83 Note: computed as $\frac{(42-27)/(70-42)}{27/42}$
- **b**) central F distribution with 28 and 42 df.
- **c**) $p > 0.05$.
- **d**) No evidence of an interaction between angle and distance effects.

7. The additional information is

- **a**) df: distance 1, angle 9, error 74
- **b**) SS: 6470. Note: computed as the difference in SSE between model (6, linear distance + angle) and model (2, angle only).

8. The requested test is a comparison of models (1) and (6).

a) F = 1.25 Note: computed as $\frac{(45-42)/(74-70)}{42/70}$.

b) central F distribution with 3 and 70 df.

c) $p > 0.05$.

d) No evidence of lack of fit for the linear distance model.

9. Yes, the error SS for model (6) will always be $\geq$ the error SS for model (1). The column space of the $\boldsymbol{X}$ matrix for model (6) is a subspace of the column space of the $\boldsymbol{X}$ matrix for model (1).

10. No evidence that model (6) is inappropriate. From the residual plot:
There are no obvious outliers, no lack of fit, and no evidence of unequal variances.
From the Cook's distance plot:
No unusually influential observations. All D_i are less than 1

11. (67.8, 69.2)
Note: Computed as $68.54 \pm t_{0.975,\,70} \times 0.36$

12. No - the concern is about a single observation. The appropriate comparison is with a prediction interval for a single value, not with a confidence interval for the mean.
Note: the prediction sd is $\sqrt{0.36^2 + 49.06/70} = 0.91$. So the prediction interval is (66.7, 70.4), which includes 70, the value of concern.

13. A reasonable model is

$$\begin{aligned} Y_{ij} &= \beta_0 + \beta_1\, d_i(x, y) + \beta_2\, \cos(a_{ij}(x, y)) + \varepsilon_{ij} \\ \varepsilon_{ij} &\overset{iid}{\sim} N(0, \sigma^2) \end{aligned} \tag{1}$$

14. Non-linear least squares
Since errors are assumed normal with equal variance, least squares is appropriate.
A description in terms of a loss function such as $\Sigma(Y_{ij} - \widehat{Y}_{ij})^2$ or a log likelihood was accepted for full credit.
Notes: Given (x, y), model (1) is linear in the parameters $(\beta_0, \beta_1, \beta_2)$. But the model is non-linear in (x, y). Such models are sometimes called partially linear models.

15. The appropriate test statistic is Hotelling's T^2:

$$\boldsymbol{T^2 = (l - \kappa)'\widehat{\Sigma}^{-1}(l - \kappa)}$$

Note: The name (Hotelling's T^2) was not required.

16. The contributions to the log likelihood are:

a)
$$\log L_{11} = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left[\frac{Y_{11} - f(x, y)}{\sigma}\right]^2$$

b)
$$\log L_{14} = \log \Phi\left[\frac{Y_{14} - f(x, y)}{\sigma}\right],$$

where Φ is the standard normal cumulative distribution function.

Part I

Consider the linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is an n-dimensional vector of response values, $\boldsymbol{X}$ is a known $n \times p$ matrix of rank r, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$ for some unknown variance parameter $\sigma^2 > 0$.

1. Suppose $\boldsymbol{c}$ is a known p-dimensional vector. What must be true of $\boldsymbol{c}$ in order for $\boldsymbol{c}'\boldsymbol{\beta}$ to be estimable?

2. Suppose $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable. Provide an expression for the best linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$ in terms of $\boldsymbol{c}$, $\boldsymbol{X}$, and $\boldsymbol{y}$.

3. Suppose $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable. Find the distribution of the best linear unbiased estimator of $\boldsymbol{c}'\boldsymbol{\beta}$.

4. Suppose $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable. Provide an expression for a 95% confidence interval for $\boldsymbol{c}'\boldsymbol{\beta}$.

5. Show that the confidence interval provided in problem **4** has coverage probability 0.95.

Part II

Researchers planted 100 maize genotypes in each of 10 fields. Each field contained 100 plots of soil. The 100 genotypes were randomly assigned to the 100 plots in each field, with 1 plot per genotype per field. After 6 seeds of the assigned genotype were planted in each plot, cameras were placed next to the plots, with one camera per plot. Each camera was programmed to take and transmit an image of the 6 plants in its plot at 9:00 A.M. each day for 50 consecutive days. The first images were taken shortly after all plants emerged from the soil. This process produced a total of 50,000 images.

The 50,000 images were partitioned into 500 sets of 100 images. The 100 images in any given set were taken by the 100 cameras in a given field on a given day. Each set of images was assigned to 2 different workers for processing, and each worker processed only 1 set of images. The workers (1,000 in total) were hired through Amazon Mechanical Turk, a system that provides access to independent workers who perform tasks remotely with a computer connected to the Internet. To process an image, workers were asked to indicate the top of each plant by pointing a computer cursor at the appropriate position within the image and clicking the computer mouse. Image processing software was used to find the bottom of each plant in each image. Given the cursor locations associated with worker mouse clicks and the results of the image processing software, the height of each plant in each image was approximated.

Problems **6** through **13** focus on the analysis of images taken on day 50. Let y_{fgpw} be the approximated height on day 50 for field f, genotype g, plant p, and worker w assigned to process images from field f on day 50 ($f = 1, \ldots, 10$, $g = 1, \ldots, 100$, $p = 1, \ldots, 6$, $w = 1, 2$). Suppose

$$\begin{aligned} y_{fgpw} &= \mu + \phi_f + \gamma_g + a_{fg} + b_{fgp} + c_{fw} + e_{fgpw}, \\ a_{fg} &\sim N(0, \sigma_a^2), \\ b_{fgp} &\sim N(0, \sigma_b^2), \\ c_{fw} &\sim N(0, \sigma_c^2), \text{ and} \\ e_{fgpw} &\sim N(0, \sigma_e^2), \end{aligned} \tag{1}$$

where μ, $\phi_1, \ldots, \phi_{10}$, $\gamma_1, \ldots, \gamma_{100}$ are unknown parameters in $\mathbb{R}$, $\sigma_a^2, \sigma_b^2, \sigma_c^2$, and σ_e^2 are unknown positive variance components, and all random effects and errors are independent.

6. Let $\bar{y}_{\cdot g \cdot\cdot}$ be the average of all the approximated plant heights on day 50 for genotype g ($g = 1, \ldots, 100$). Find the expected value of $\bar{y}_{\cdot 1 \cdot\cdot}$ in terms of Model (1) parameters.

7. Find the variance of $\bar{y}_{\cdot 1 \cdot\cdot}$ in terms of Model (1) parameters.

8. Find the variance of $\bar{y}_{\cdot 1 \cdot\cdot} - \bar{y}_{\cdot 2 \cdot\cdot}$ in terms of Model (1) parameters.

9. Model (1) was fit to the approximated plant heights for day 50. A portion of the `R` code and output is provided on page 5. Determine the value of a t-statistic for testing $H_0 : \gamma_9 = \gamma_{10}$.

10. Determine the degrees of freedom for the t-statistic in problem **9**.

11. The REML method was used to estimate the variance components in Model (1). The REML method involves a likelihood function for observations known as error contrasts. How many error contrasts are involved in the REML likelihood function for the fit of Model (1) to the approximated heights collected for day 50?

12. Suppose some workers tend to position the cursor higher than they should while other workers tend to position the cursor lower than they should when indicating the tops of plants in images. Explain either how this type of variability is accounted for in Model (1) or how to modify Model (1) to account for such variability.

13. Suppose some workers are more careless than others when indicating the tops of plants in images. Careless workers may position the cursor far too high for some plants and far too low for others. Explain either how this type of variability among workers is accounted for in Model (1) or how to modify Model (1) to account for such variability.

Part III

For any given field, genotype, and day, 12 approximated heights are included in the dataset because each of 2 workers provides height approximations for each of 6 plants. Let $\bar{y}_{fgd}$ be the average of the 12 approximated heights for field f, genotype g, and day d ($f = 1, \ldots, 10$, $g = 1, \ldots, 100$, $d = 1, \ldots, 50$). Suppose

$$\bar{y}_{fgd} = \bar{\mu} + \bar{\phi}_f + \bar{\gamma}_g + u_{1g}q_1(d) + u_{2g}q_2(d) + u_{3g}q_3(d) + v_{1fg}r_1(d) + v_{2fg}r_2(d) + \bar{e}_{fgd}, \quad (2)$$

$$u_{1g} \sim N(0, \sigma^2_{u1}),\ u_{2g} \sim N(0, \sigma^2_{u2}), u_{3g} \sim N(0, \sigma^2_{u3}),$$

$$v_{1fg} \sim N(0, \sigma^2_{v1}), v_{2fg} \sim N(0, \sigma^2_{v2}),\ \text{and } \bar{e}_{fgd} \sim N(0, \bar{\sigma}^2_e),$$

where $\bar{\mu}$, $\bar{\phi}_1, \ldots, \bar{\phi}_{10}$, and $\bar{\gamma}_1, \ldots, \bar{\gamma}_{100}$ are unknown parameters in $\mathbb{R}$, σ^2_{u1}, σ^2_{u2}, σ^2_{u3}, σ^2_{v1}, σ^2_{v2}, and $\bar{\sigma}^2_e$ are unknown positive variance components, $q_1(\cdot)$, $q_2(\cdot)$, $q_3(\cdot)$, $r_1(\cdot)$, and $r_2(\cdot)$ are fixed known functions that map $\mathbb{R}$ to $\mathbb{R}$, and all random terms are independent. [*Note that you are not expected to specify the functions $q_1(\cdot)$, $q_2(\cdot)$, $q_3(\cdot)$, $r_1(\cdot)$, and $r_2(\cdot)$. You may assume these are fixed known functions specified by the researchers.*]

14. In terms of Model (2) parameters, find an expression for the correlation between average approximated height on day 10 and the average approximated height on day 20 for plants of genotype 1 in field 1.

15. Suppose the researchers ask for "the estimated growth curve for genotype 1." Assuming Model (2) holds, explain how you would use results from the fit of Model (2) to provide a growth curve that addresses the researchers' request.

Part IV

Shortly after day 50, the researchers exposed all 6,000 plants to a pathogen. The field cameras were programmed to take and transmit an image of the plants two weeks after pathogen exposure. The 1,000 images were examined by an expert who indicated the presence or absence of pathogen damage for each of the 6,000 plants. Let z_{fgp} be 1 if plant p of genotype g in field f sustained pathogen damage according to the expert's evaluation and 0 otherwise ($f = 1, \ldots, 10$, $g = 1, \ldots, 100$, $p = 1, \ldots, 6$). Suppose for some parameters $\pi_1, \ldots, \pi_{100} \in (0, 1)$,

$$z_{fgp} \sim \text{Bernoulli}(\pi_g), \text{ and all } z_{fgp} \text{ terms are independent.} \quad (3)$$

16. Suppose $\sum_{f=1}^{10} \sum_{p=1}^{6} z_{f1p} = 23$. Find a confidence interval for π_1 with confidence level approximately equal to 0.95.

17. The number of plants with pathogen damage per plot for 10 fields and a subset of genotypes is provided below in Table 1. Based on your review of the data in Table 1, provide a reason why Model (3) may be inappropriate.

18. Propose a model for the full dataset $\{x_{fg} \equiv \sum_{p=1}^{6} z_{fgp} : f = 1, \ldots, 10,\ g = 1, \ldots, 100\}$ that you would fit to gain insights the researchers might find interesting.

Table 1. Number of plants with pathogen damage by field and genotype.

	Genotype							
Field	1	2	3	4	5	$\cdots$	100	Total
1	6	0	6	6	0	$\cdots$	5	427
2	0	0	6	0	0	$\cdots$	4	155
3	6	0	6	0	0	$\cdots$	0	219
4	0	5	6	6	0	$\cdots$	1	490
5	0	0	6	0	0	$\cdots$	6	19
6	0	3	6	0	0	$\cdots$	0	51
7	0	1	6	0	0	$\cdots$	6	366
8	6	0	6	4	0	$\cdots$	0	74
9	0	0	6	0	0	$\cdots$	0	59
10	5	1	6	1	0	$\cdots$	4	292
Total	23	10	60	17	0	$\cdots$	26	

Partial R Code and Output for the Fit of Model (1).

```
> o = lmer(y ~ field + genotype + (1 | a) + (1 | b) + (1 | c))

> summary(o)

Linear mixed model fit by REML ['lmerMod']
Formula: y ~ field + genotype + (1 | a) + (1 | b) + (1 | c)

REML criterion at convergence: 51937.8




Fixed effects:
             Estimate Std. Error t value
(Intercept) 152.35029    1.09520 139.107
field2       -8.91885    1.41863  -6.287
field3       -6.63456    1.41863  -4.677
field4        0.06360    1.41863   0.045
field5        5.22145    1.41863   3.681
field6       -1.62157    1.41863  -1.143
field7       -6.10325    1.41863  -4.302
field8       13.51825    1.41863   9.529
field9        2.24084    1.41863   1.580
field10       2.32538    1.41863   1.639
genotype2    13.01750    0.62477  20.836
genotype3    -4.02125    0.62477  -6.436
genotype4     6.06067    0.62477   9.701
genotype5    23.39275    0.62477  37.442
genotype6     0.38508    0.62477   0.616
genotype7   -10.55892    0.62477 -16.901
genotype8    -6.23683    0.62477  -9.983
genotype9    11.99942    0.62477  19.206
genotype10   -4.50458    0.62477  -7.210
[90 ADDITIONAL LINES OMITTED]
```

1. The linear combination $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable if and only if $\boldsymbol{c} = \boldsymbol{X}'\boldsymbol{a}$ for some n-dimensional vector $\boldsymbol{a}$.

2. The BLUE of $\boldsymbol{c}'\boldsymbol{\beta}$ is $\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$.

3. Because linear transformations of multivariate normal vectors are also multivariate normal, $\boldsymbol{y}$ is multivariate normal as a linear transformation of $\boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Because $\boldsymbol{y}$ is multivariate normal and the BLUE $\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$ is a linear transformation of $\boldsymbol{y}$, it follows that the BLUE is multivariate normal. To determine the mean and variance of the BLUE, let $\boldsymbol{a}$ be such that $\boldsymbol{c}' = \boldsymbol{a}'\boldsymbol{X}$ so that $\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{a}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}$. Then note that

$$\begin{aligned}
\mathrm{E}(\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}) &= \mathrm{E}(\boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\mathrm{E}(\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\mathrm{E}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\
&= \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}(\boldsymbol{X}\boldsymbol{\beta} + \mathrm{E}(\boldsymbol{\epsilon})) = \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{a}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{c}'\boldsymbol{\beta},
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Var}(\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}) &= \mathrm{Var}(\boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}) = \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\mathrm{Var}(\boldsymbol{y})\boldsymbol{P}'_{\boldsymbol{X}}\boldsymbol{a} = \boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}(\sigma^2\boldsymbol{I})\boldsymbol{P}'_{\boldsymbol{X}}\boldsymbol{a} \\
&= \sigma^2\boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}'_{\boldsymbol{X}}\boldsymbol{a} = \sigma^2\boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{a} = \sigma^2\boldsymbol{a}'\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{a} \\
&= \sigma^2\boldsymbol{a}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{a} = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{c}.
\end{aligned}$$

Thus, $\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y} \sim N(\boldsymbol{c}'\boldsymbol{\beta}, \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{c})$.

4. Let $\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{y}$. Let $t^* = t_{n-r,0.975}$ be the 0.975 quantile of the central t distribution with $n - r$ degrees freedom. Let $s^2 = \boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}/(n - r)$. Then a 95% confidence interval for $\boldsymbol{c}'\boldsymbol{\beta}$ is

$$\left[\boldsymbol{c}'\hat{\boldsymbol{\beta}} - t^*\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{c}}, \boldsymbol{c}'\hat{\boldsymbol{\beta}} + t^*\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{c}}\right].$$

5. First, note that

$$z \equiv \frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}} - \boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{c}}} \sim N(0, 1)$$

by the solution to problem 3. Next, note that

$$w^2 \equiv \frac{(n-r)s^2}{\sigma^2} = \boldsymbol{y}'\left(\frac{\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}}{\sigma^2}\right)\boldsymbol{y} \sim \chi^2_{n-r}$$

because

$$\left(\frac{\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}}{\sigma^2}\right)\mathrm{Var}(\boldsymbol{y}) = \boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}$$

is an idempotent matrix, $\left(\frac{\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}}}{\sigma^2}\right)$ is of rank $n - r$, and

$$[\mathrm{E}(\boldsymbol{y})]'\left(\frac{\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}}}{\sigma^2}\right)\mathrm{E}(\boldsymbol{y}) = \frac{1}{\sigma^2}\boldsymbol{\beta}'\boldsymbol{X}'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X}\boldsymbol{\beta} = 0.$$

Next, note that $\boldsymbol{c}'\hat{\boldsymbol{\beta}} = \boldsymbol{a}'\boldsymbol{P_X}\boldsymbol{y}$ and $w^2 = \frac{(n-r)s^2}{\sigma^2} = \boldsymbol{y}'\left(\frac{\boldsymbol{I}-\boldsymbol{P_X}}{\sigma^2}\right)\boldsymbol{y}$ are independent because

$$\boldsymbol{a}'\boldsymbol{P_X}(\sigma^2\boldsymbol{I})\left(\frac{\boldsymbol{I}-\boldsymbol{P_X}}{\sigma^2}\right) = \boldsymbol{a}'\boldsymbol{P_X}(\boldsymbol{I}-\boldsymbol{P_X}) = \boldsymbol{0}'.$$

It follows that $\boldsymbol{z} = \frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}}-\boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}}$ and $w^2 = \frac{(n-r)s^2}{\sigma^2}$ are independent, and thus,

$$\begin{aligned}
\frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}}-\boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}} &= \frac{(\boldsymbol{c}'\hat{\boldsymbol{\beta}}-\boldsymbol{c}'\boldsymbol{\beta})/\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}/\sqrt{\sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}} \\
&= \frac{z}{\sqrt{[(n-r)s^2/\sigma^2]/(n-r)}} = \frac{z}{\sqrt{w^2/(n-r)}} \sim t_{n-r}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left\{\boldsymbol{c}'\boldsymbol{\beta} \in \left[\boldsymbol{c}'\hat{\boldsymbol{\beta}} - t^*\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}, \boldsymbol{c}'\hat{\boldsymbol{\beta}} + t^*\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}\right]\right\} \\
&= \mathbb{P}\left\{\boldsymbol{c}'\hat{\boldsymbol{\beta}} - t^*\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}} \leq \boldsymbol{c}'\boldsymbol{\beta} \leq \boldsymbol{c}'\hat{\boldsymbol{\beta}} + t^*\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}\right\} \\
&= \mathbb{P}\left\{-t^* \leq \frac{\boldsymbol{c}'\boldsymbol{\beta}-\boldsymbol{c}'\hat{\boldsymbol{\beta}}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}} \leq t^*\right\} \\
&= \mathbb{P}\left\{t^* \geq \frac{\boldsymbol{c}'\hat{\boldsymbol{\beta}}-\boldsymbol{c}'\boldsymbol{\beta}}{\sqrt{s^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}}} \geq -t^*\right\} = 0.95.
\end{aligned}$$

6. First, note that

$$\bar{y}_{.1..} = \mu + \bar{\phi}_{.} + \gamma_1 + \bar{a}_{.1} + \bar{b}_{.1.} + \bar{c}_{..} + \bar{e}_{.1..}$$

Because all the random terms have mean 0, $\mathrm{E}(\bar{y}_{.1..}) = \mu + \bar{\phi}_{.} + \gamma_1$

7.

$$\begin{aligned}
\mathrm{Var}(\bar{y}_{.1..}) &= \mathrm{Var}(\mu + \bar{\phi}_{.} + \gamma_1 + \bar{a}_{.1} + \bar{b}_{.1.} + \bar{c}_{..} + \bar{e}_{.1..}) \\
&= \mathrm{Var}(\bar{a}_{.1} + \bar{b}_{.1.} + \bar{c}_{..} + \bar{e}_{.1..}) \\
&= \mathrm{Var}(\bar{a}_{.1}) + \mathrm{Var}(\bar{b}_{.1.}) + \mathrm{Var}(\bar{c}_{..}) + \mathrm{Var}(\bar{e}_{.1..}) \\
&= \sigma_a^2/10 + \sigma_b^2/60 + \sigma_c^2/20 + \sigma_e^2/120
\end{aligned}$$

8.

$$\begin{aligned}
\mathrm{Var}(\bar{y}_{.1..} - \bar{y}_{.2..}) &= \mathrm{Var}(\bar{a}_{.1} - \bar{a}_{.2} + \bar{b}_{.1.} - \bar{b}_{.2.} + \bar{e}_{.1..} - \bar{e}_{.2..}) \\
&= \sigma_a^2/5 + \sigma_b^2/30 + \sigma_e^2/60
\end{aligned}$$

9. Because of the balanced experimental design, the standard error for $\widehat{\gamma_i - \gamma_j}$ is the same for all $i \neq j$. The R output shows 0.62477 is the standard error for $\widehat{\gamma_i - \gamma_1}$ for $i = 2, \ldots, 10$. Thus, the standard error for $\widehat{\gamma_9 - \gamma_{10}}$ is also 0.62477, and $t = [11.99942 - (-4.50458)]/0.62477 \approx 26.4$.

10. The degrees of freedom should be $(10-1)(100-1) = 891$, which matches the degrees of freedom for error in a randomized complete block design with 10 blocks and 100 treatments. To see why, note that we have a balanced design with multiple observations per experimental unit, where plots are experimental units. Inferences obtained from the linear mixed-effects model analysis of the full day-50 dataset will match inferences obtained from an analysis of experimental unit averages. The model for experimental unit averages implied by the model for the full day-50 dataset is

$$\begin{aligned} \bar{y}_{fg\cdot\cdot} &= \mu + \phi_f + \gamma_g + a_{fg} + \bar{b}_{fg\cdot} + \bar{c}_{f\cdot} + \bar{e}_{fg\cdot\cdot} \\ &= \mu + (\phi_f + \bar{c}_{f\cdot}) + \gamma_g + (a_{fg} + \bar{b}_{fg\cdot} + \bar{e}_{fg\cdot\cdot}) \\ &= \mu + (\phi_f + \bar{c}_{f\cdot}) + \gamma_g + (a_{fg} + \bar{b}_{fg\cdot} + \bar{e}_{fg\cdot\cdot}) \\ &= \mu + \beta_f + \gamma_g + \epsilon_{fg}, \end{aligned}$$

where $\beta_f \equiv \phi_f + \bar{c}_{f\cdot}$ and $\epsilon_{fg} \equiv a_{fg} + \bar{b}_{fg\cdot} + \bar{e}_{fg\cdot\cdot}$. Because the ϵ_{fg} error terms are independent and identically distributed normal random variables with mean zero, we can analyze the experimental unit averages with the usual linear model for a randomized complete block design.

11. $n - \text{rank}(\boldsymbol{X}) = 12000 - 109 = 11891$

12. The c_{fw} random effects are intended to account for this type of variation among workers.

13. This type of variability among workers is not accounted for in the model. Rather than assuming σ_e^2 is the variance of each error term, we could consider allowing the error variance to be different for each worker (i.e., $\text{Var}(e_{fgpw}) = \sigma_{efw}^2$ for each combination of $f = 1, \ldots, 10$ and $w = 1, 2$).

14. First, note that, for any $d \in \{1, \ldots, 50\}$,

$$\begin{aligned} \text{Var}(\bar{y}_{fgd}) &= \text{Var}(u_{1g}q_1(d) + u_{2g}q_2(d) + u_{3g}q_3(d) + v_{1fg}r_1(d) + v_{2fg}r_2(d) + \bar{e}_{fgd}) \\ &= \sum_{i=1}^{3} \sigma_{ui}^2 q_i^2(d) + \sum_{j=1}^{2} \sigma_{vj}^2 r_j^2(d) + \bar{\sigma}_e^2 \end{aligned}$$

Also, for any field and genotype, the covariance for days 10 and 20 is

$$\begin{aligned} &\text{Cov}\left[\sum_{i=1}^{3} u_{ig}q_i(10) + \sum_{j=1}^{2} v_{jg}r_j(10), \sum_{i=1}^{3} u_{ig}q_i(20) + \sum_{j=1}^{2} v_{jg}r_j(20)\right] \\ &= \sum_{i=1}^{3} \sigma_{ui}^2 q_i(10)q_i(20) + \sum_{j=1}^{2} \sigma_{vj}^2 r_j(10)r_j(20) \end{aligned}$$

Thus, the correlation is

$$\frac{\sum_{i=1}^{3}\sigma_{ui}^2 q_i(10)q_i(20)+\sum_{j=1}^{2}\sigma_{vj}^2 r_j(10)r_j(20)}{\sqrt{\left(\sum_{i=1}^{3}\sigma_{ui}^2 q_i^2(10)+\sum_{j=1}^{2}\sigma_{vj}^2 r_j^2(10)+\bar{\sigma}_e^2\right)\left(\sum_{i=1}^{3}\sigma_{ui}^2 q_i^2(20)+\sum_{j=1}^{2}\sigma_{vj}^2 r_j^2(20)+\bar{\sigma}_e^2\right)}}$$

15. The empirical best linear unbiased predictor

$$\hat{\bar{\mu}}+\frac{1}{10}\sum_{f=1}^{10}\hat{\bar{\phi}}_f+\hat{\bar{\gamma}}_1+\sum_{i=1}^{3}\hat{u}_{i1}q_i(d)$$

for $d\in[1,50]$ provides the most relevant estimator of "the growth curve for genotype 1," where $\hat{\bar{\mu}}+\frac{1}{10}\sum_{f=1}^{10}\hat{\bar{\phi}}_f+\hat{\bar{\gamma}}_1$ is the generalized least squares estimator of $\bar{\mu}+\frac{1}{10}\sum_{f=1}^{10}\bar{\phi}_f+\bar{\gamma}_1$ and $\hat{u}_{11},\hat{u}_{21}$, and $\hat{u}_{31}$ are the empirical BLUPs of u_{11},u_{21}, and u_{31}, respectively.

16. It is straightforward to show that $\hat{\pi}_1=23/60$ is the MLE of π_1 and that an evaluation of inverse Fisher information leads to $\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/60}$ as a standard error. Thus, an appropriate Wald interval is

$$\left[\hat{\pi}_1-1.96\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/60},\hat{\pi}_1+1.96\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/60}\right]\approx[0.26,0.51]$$

17. If Model (3) were correct, the 23 plants of gentoype 1 damaged by pathogen would be expected to be more evenly dispersed over the 10 fields rather than concentrated on just 4 of the 10 fields. There seem to be clear field effects (e.g., see large variation in Table 1 row totals), indicating that the probability of damage from the pathogen is not constant across fields within genotypes. Also, there may be field $\times$ genotype interactions (corresponding to random effects of plots in fields, for example). Evidence for such interactions can be informally noted by the lack of correlation between field totals and genotype-specific field totals.

18. Multiple answers are possible. To avoid fitting problems that will occur because at least one genotype has no damaged plants (and another has all damaged plants), the following random-effects model may be useful. For $f=1,\ldots,10$ and $g=1,\ldots,100$, suppose

$$x_{fg}\sim\text{Bernoulli}(\theta_{fg}),\ \log\left(\frac{\theta_{fg}}{1-\theta_{fg}}\right)=\mu+a_f+b_g+c_{fg},$$

where μ is an unknown parameter, the a_f terms are $N(0,\sigma_a^2)$, the b_g terms are $N(0,\sigma_b^2)$, the c_{fg} terms are $N(0,\sigma_c^2)$, σ_a^2, σ_b^2, and σ_c^2 are unknown positive variance components, all the random effects are independent, and the x_{fg} terms are conditionally independent given the random effects. The researchers would likely be interested in the empirical BLUPs of the random effects $b_1,\ldots,b_{100}$ as these could be used to rank genotypes with respect to pathogen susceptibility.

Meta-analysis can be defined as a systematic approach to the summarization and synthesis of existing information about a given problem. An area in which meta-analysis is widely used is the field of medicine when multiple clinical studies have been published about treatments for some disease or condition. In this context, a meta-analysis generally involves searching the literature for papers that have been published on the topic of interest, narrowing the collection of studies obtained through the application of certain criteria for inclusion in the meta-analysis, and producing a combined estimate of an overall treatment effect. Although the entire process involves a number of steps that are not statistical in nature, the production of an overall estimate of a treatment effect, sometimes called meta-analytic estimation, is very much a statistical topic.

To set the stage for the estimation problem, suppose that in a collection of studies we can obtain, from each individual study, an estimate of a treatment effect T_j and an estimate of the variance of T_j, V_j, for $j = 1, \ldots, K$. Exactly what type of statistics the T_j are will depend on the problem under investigation, and we will consider this in more detail when needed in what follows. For now, we can suppose that the T_j might be differences between mean responses in treatment groups and mean responses in control groups. The variances V_j are typically the squares of standard errors. The important point is that, as estimators, the T_j and the V_j are random variables.

There are two models used extensively in meta-analytic estimation, one called the *fixed effect* model and the other called the *random effects* model. We will be concerned here only with the random effects model. This model was originally introduced by Rebecca DerSimonian and Nan Laird in 1986, "Meta-Analysis in Clinical Trials", *Controlled Clinical Trials* **7**, 177-188. For $j = 1, \ldots, K$, the model has the form

$$T_j \overset{ind}{\sim} \mathrm{N}(\theta_j, V_j) \quad \text{and} \quad \theta_j \overset{iid}{\sim} \mathrm{N}(\mu, \tau^2) \tag{1}$$

In this model, the V_j are assumed to be fixed and known with values equal to those reported in the individual studies, leaving only two unknown parameters to be estimated: μ and τ^2.

1. Derive the MLE for μ and τ^2 in model (1).

2. Using either the observed or expected information matrix, give quantities that you would use as standard errors to compute Wald confidence intervals for μ and τ^2.

Many statisticians would find it natural to consider a Bayesian analysis of model (1).

3. Consider using a joint prior in product form, $\pi(\mu, \tau^2) = \pi_1(\mu)\,\pi_2(\tau^2)$, or, equivalently, $\pi(\mu, \tau) = \pi_1(\mu)\,\pi_2(\tau)$ where the prior for μ is normal with mean M_0 and variance V_0, and the prior for τ (not τ^2) is Unif(0, A) with $A < \infty$. A number of MCMC algorithms could be used to simulate values from the joint posterior of μ and τ^2. Suppose we have decided to use an overall Gibbs sampling structure for our algorithm. We have the choice of using the joint posterior

$$p(\mu, \tau^2|\boldsymbol{t}) \propto h(\boldsymbol{t}|\mu, \tau^2)\,\pi_1(\mu)\,\pi_2(\tau^2) = \left[\prod_{j=1}^{K} h(t_j|\mu, \tau^2)\right] \pi_1(\mu)\,\pi_2(\tau^2)$$

or the joint posterior

$$\begin{aligned} p(\mu, \tau^2, \boldsymbol{\theta}|\boldsymbol{t}) &\propto f(\boldsymbol{t}|\boldsymbol{\theta})\,g(\boldsymbol{\theta}|\mu, \tau^2)\,\pi_1(\mu)\,\pi_2(\tau^2) \\ &\propto \left[\prod_{j=1}^{K} f_j(t_j|\theta_j)\right] \left[\prod_{j=1}^{K} g(\theta_j|\mu, \tau^2)\right] \pi_1(\mu)\,\pi_2(\tau^2) \end{aligned}$$

Using either of these forms, the full conditional posterior of μ will be normal in form. Derive the full conditional posterior of τ^2 for each of the two joint posterior forms. *Note: For the joint posterior form given as $p(\mu, \tau^2, \boldsymbol{\theta}|\boldsymbol{t})$ do not worry about sampling from the full conditional posteriors of the θ_j as these will again be normal in form.*

4. Using whichever full conditional posterior of τ^2 you believe would result in a simpler overall Gibbs Sampling algorithm, give a simple sub-algorithm for sampling from the full conditional posterior of τ^2 in this algorithm. That is, given that you have current value $\mu^{(m)}$ in the first form, or values $\mu^{(m)}$ and $\theta_1^{(m)}, \theta_2^{(m)}, \ldots, \theta_K^{(m)}$ in the second form, give a very short algorithm to sample one value from $p(\tau^2|\cdot)$.

Suppose that, rather than a difference in treatment group means, the values T_j were observed proportions for the prevalence (or incidence) of some disease. For example, a recent paper by Kalilani, Sun, Pelgrims, Noack-Rink and Billanueva (2018) conducted a meta-analysis using 36 studies that reported the number of cases of drug-resistant epilepsy out of samples of populations of epilepsy patients. Here, each study reported the sample size n_j and the number of drug resistant cases Y_j, so the estimated "effects" in this case were

$$T_j = Y_j/n_j$$

and the estimated individual study variances could be computed as

$$V_j = \frac{T_j\,(1 - T_j)}{n_j}.$$

A plot of 95% approximate confidence intervals for the true proportion in each study is shown in Figure 1.

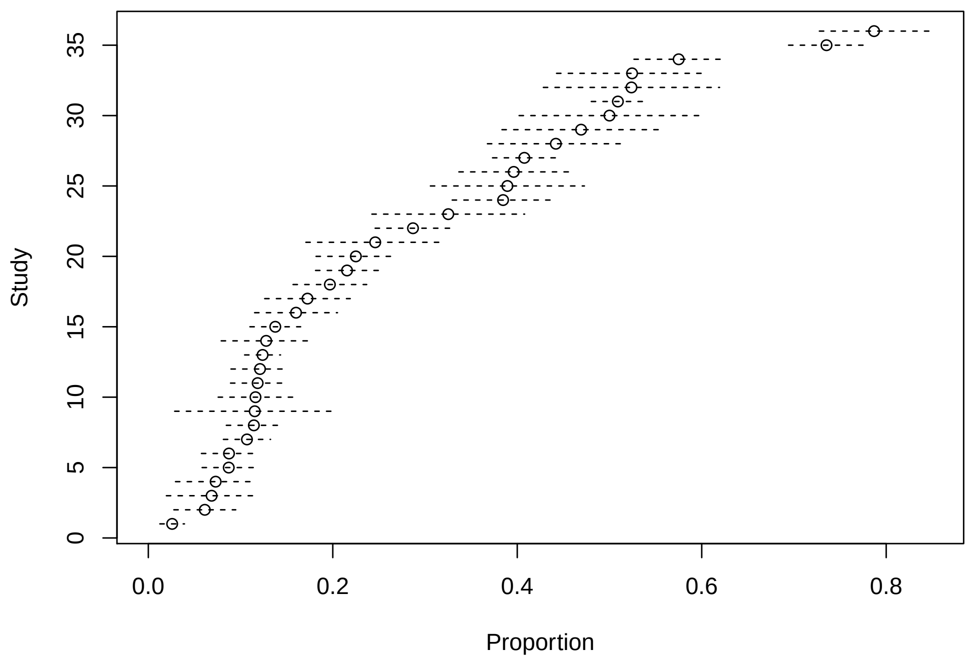

Figure 1: Plot of interval estimates, each with coverage level approximately 95%, for the proportion of drug resistant epilepsy from 36 individual studies.

Suppose we would like to use model (1) in a meta-analysis of these studies. One might hope to motivate the normal model for the T_j in this case, as well as the form of the V_j, by asymptotic normality of maximum likelihood estimators and inverse expected information (the Central Limit Theorem might also suffice). This theory was, after all, used in construction of the intervals presented in Figure 1. We might be curious about how large the n_j need to be before the asymptotic sampling distribution provides reasonable motivation for the normal-based model (1).

To examine the adequacy of this normal approximation, we might conduct a simple Monte Carlo investigation, using randomly generated values from a binomial distribution with parameter $p = 0.10$ and sample sizes of $n = 50, 100, 500,$ and 1000. To formalize this exercise, let y_r denote the binomial value (as a count) generated during Monte Carlo replicate $r = 1, \ldots, R$. Define

$$\hat{p}_r = \frac{y_r}{n} \qquad \text{and} \qquad \hat{v}(\hat{p}_r) = \frac{\hat{p}_r\,(1 - \hat{p}_r)}{n}.$$

Suppose we decide to examine the limiting normal result through the use of a Kolmogorov test based on the R values of $\hat{p}_r$ and of $\hat{v}(\hat{p}_r)$. Conducting this exercise for values of $n \in \{50,\ 100,\ 500,\ 1000\}$ and $R \in \{500,\ 1000,\ 5000\}$ produced the values given in Table 1.

Table 1: Kolmogorov p-value for tests of normality and coverage of 95% two-sided Wald confidence intervals for R Monte Carlo simulations for binomial distributions with sample size n and success probability $p = 0.10$.

n	R	p-value	coverage (%)
50	500	8.6×10^{-6}	84.8
	1000	$< 2 \times 10^{-16}$	87.4
	5000	$< 2 \times 10^{-16}$	88.1
100	500	0.0002	93.0
	1000	4.1×10^{-10}	93.6
	5000	$< 2 \times 10^{-16}$	93.1
500	500	0.039	95.8
	1000	2.1×10^{-6}	94.6
	5000	2.7×10^{-7}	94.3
1000	500	0.341	95.6
	1000	0.231	95.1
	5000	0.0006	95.2

5. The issue of whether Wald intervals for small values of estimated binomial parameters are contained entirely within the parameter space is a different issue than asymptotic normality. To verify this, fix a binomial random variable at a value of $y = 3$ and determine the range of values of binomial sample size n that will then result in a 95% two-sided Wald interval that does not extend below 0.

6. Give the quantities that were tested for normality in the Monte Carlo exercise reported in Table 1, and state the null hypothesis of the test. Note that $\hat{p}_r$ and $\hat{v}(\hat{p}_r)$ are functions of n.

Hint: Consider carefully exactly what quantities are technically converging in distribution in the result that allows us to say that $\hat{p}_n$ is asymptotically normal.

7. In Table 1, there seems to be some discrepancy in results of the test for normality (the p-values in the table) between situations with smaller values of R and those with larger values of R for some n such as $n = 500$ and $n = 1000$. Explain why this occurs, and how that might impact the issue of whether model (1) can be motivated by asymptotic normality arguments for binomial proportions.

Suppose that we are willing to accept asymptotic normality of $T_j = Y_j/n_j$ as motivation for the normal-based model (1) in our meta-analysis (regardless of what you might think of that in light of Table 1, just suppose we accept it). So we have the T_j being normally distributed with expectations θ_j and the θ_j as normally distributed with common expected value μ. Continue to consider a Bayesian approach to estimation, as begun in Problems **3** and **4**.

8. We will again use a joint prior formulated in product form (as in Problem **3**). Comment on whether it is still reasonable to take the prior for μ to be normal with mean M_0 and variance V_0. Specifically, the normal model form has been motivated by asymptotic normality of the T_j. Does this then also motivate the use of a normal prior for μ, perhaps because it is still a location parameter? Are there any potential difficulties that might arise from the use of a normal prior for μ? If so, do you see any clear alternatives to a normal prior?

9. Similar to issues raised in Problem **8**, is it still reasonable to take a prior for τ^2 to result from a proper uniform distribution for τ on a (presumably wide) interval $(0, A)$?

At this point, return to consideration of a general problem in which we have no concerns about modeling the T_j from individual studies as normal random variables with means θ_j and the θ_j as normal random variables with mean μ. What we wish to examine now about model (1) is the assumption that the V_j can be taken as known true variances for the individual estimated T_j. In applications, these are also estimated values and we may wish to consider how we could incorporate uncertainty about the true variances of the T_j in the model. It would seem immediate that we could consider modeling both the T_j and the V_j in the data model. We will consider three options for doing so, all building on the motivation used in model (1) for the T_j that the data model mimic to some degree the approximate sampling distribution of the T_j. To reduce the overall complexity of our current topic, consider the simple case in which the T_j are sample means of independent and identically distributed random variables.

Option 1. As a first possibility consider modeling T_j and V_j from individual studies as normal variates, which might be motivated by joint asymptotic normality of these quantities. For this, we might take

$$\begin{aligned} T_j &\overset{ind}{\sim} N(\theta_j,\, \sigma^2) \\ V_j &\overset{iid}{\sim} N(\sigma^2,\, \psi^2) \\ \theta_j &\overset{iid}{\sim} N(\mu,\, \tau^2). \end{aligned}$$

Option 2. Now consider modifying the role of the V_j so that they are no longer variances of the T_j, but are estimated variances of the same variables for which T_j is the sample mean. Then $\hat{v}(T_j) = V_j/n_j$ where n_j is the sample size involved in study $j = 1, \ldots, K$. Asymptotic normality of $(T_j,\, V_j)$ might then motivate

$$\begin{aligned} T_j &\overset{ind}{\sim} N(\theta_j,\, \sigma^2/n_j) \\ V_j &\overset{ind}{\sim} N(\sigma^2,\, 2\sigma^4/n_j) \\ \theta_j &\overset{iid}{\sim} N(\mu,\, \tau^2). \end{aligned}$$

Option 3. Extending Option 2 to allow different individual studies to have different (true) variances results in

$$\begin{aligned} T_j &\overset{ind}{\sim} N(\theta_j,\, \sigma_j^2/n_j) \\ V_j &\overset{ind}{\sim} N(\sigma_j^2,\, 2\sigma_j^4/n_j) \\ \theta_j &\overset{iid}{\sim} N(\mu,\, \tau^2) \\ \sigma_j^2 &\overset{iid}{\sim} \mathrm{IG}(\alpha,\, \beta), \end{aligned}$$

where $\mathrm{IG}(\alpha,\, \beta)$ denotes an inverse gamma distribution with parameters α and β.

10. What desireable characteristic of maximum likelihood estimation of μ in model (1) would be lost if we used Option 1? *Hint: Examine the mle of μ you derived in Problem **1**.*

11. For the model in Option 1, we might assign σ^2 an inverse gamma prior distribution based on the mathematical convenience that results from conditional conjugacy in deriving the full conditional posterior

$$p(\sigma^2|\cdot) \propto \pi(\sigma^2) \prod_{j=1}^{K} f(t_j|\theta_j, \sigma^2),$$

which is again an inverse gamma density. Would this same motivation apply to the model in Option 2? For Option 2, show that the full conditional posterior $p(\sigma^2|\cdot)$ either is or is not an inverse gamma density under the use of an inverse gamma prior for σ^2.

12. Assume that, after assigning suitable prior distributions, we are able to successfully simulate from the joint posterior resulting from Option 2. Outline a procedure we might use to determine whether there is any motivation to consider the more complex model in Option 3.

Note: You should not need to derive explicit forms for any of the distributions you might want to use for this problem. That is, you might want to make use of $p(\mu|\boldsymbol{t}, \boldsymbol{v})$ in your answer, but you will not need to know what the formula is for this density. Also note that you will not need to derive anything based the model in Option 3. The model in Option 3 simply plays the role of the alternative under consideration.

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. Note that the model of expression (1) in the question may be written as

$$T_j = \theta_j + \epsilon_j,$$

where $\epsilon_j \sim iidN(0, V_j)$ and $\theta_j \sim iidN(\mu, \tau^2)$. Since the T_j are linear combinations of normally distributed random variables they have normal distributions. The expected values and variances are $E(T_j) = \mu$ and $\text{var}(T_j) = \tau^2 + V_j$. Since the T_j are independent, the log likelihood for μ and τ^2 may be written as

$$\begin{aligned} \ell(\mu, \tau^2) &= \sum_{j=1}^{k} \ell_j(\mu, \tau^2), \\ \ell_j(\mu, \tau^2) &= -\frac{1}{2}\sum_{j=1}^{K} \log\left(\tau^2 + V_j\right) - \frac{1}{2}\sum_{j=1}^{K}\left(\frac{(t_j-\mu)^2}{\tau^2+V_j}\right). \end{aligned}$$

Taking the derivative with respect to μ and setting it equal to zero results in

$$\hat{\mu} = \frac{\sum_{j=1}^{K}\left(\frac{t_j}{\tau^2+V_j}\right)}{\sum_{j=1}^{K}\left(\frac{1}{\tau^2+V_j}\right)}. \tag{1}$$

Taking the derivative with respect to τ^2 and setting it equal to zero results in the relation

$$\sum_{j=1}^{K}\frac{1}{\tau^2+V_j} = \sum_{j=1}^{K}\left(\frac{t_j-\mu}{\tau^2+V_j}\right)^2, \tag{2}$$

We could substitute $\hat{\mu}$ into (2), resulting in a one-dimensional maximization problem that could be solved by any number of simple algorithms (such as equal interval search) and the result substituted for τ^2 in (1). This is a version of profile likelihood.

Question 2. Taking second derivatives,

$$\frac{\partial^2}{\partial\mu^2}\ell(\mu, \tau^2) = -\sum_{j=1}^{K}\frac{1}{\tau^2+V_j}$$

$$\frac{\partial^2}{\partial\mu\,\partial\tau^2}\ell(\mu,\,\tau^2) = -\sum_{j=1}^{K}\frac{(t_j-\mu)}{(\tau^2+V_j)^2}$$

$$\frac{\partial^2}{\partial(\tau^2)^2}\ell(\mu,\,\tau^2) = \frac{1}{2}\sum_{j=1}^{K}\frac{1}{(\tau^2+V_j)^2} - \sum_{j=1}^{K}\frac{(t_j-\mu)^2}{(\tau^2+V_j)^3}$$

Taking negative expected values simplifies these expressions in that the second partial wrt μ simply loses the negative sign (but is otherwise constant), the second partial for τ^2 becomes the negative of the first term alone, and the mixed partial ends up being equal to 0 (in expectation). If we then define

$$w_j = \frac{1}{\tau^2+V_j}$$

the expected (or Fisher) information matrix becomes

$$I(\mu,\tau^2) = \begin{pmatrix} \sum w_j & 0 \\ 0 & \frac{1}{2}\sum w_j^2 \end{pmatrix}$$

The inverse of this matrix is just the reciprocal of the diagnoal elements (because the off-diagonal elements are 0) and Wald theory intervals result from asymptotic normality of $\hat{\mu}$ and $\hat{\tau}^2$ using these as the variances.

Question 3.

For use of the form $p(\mu,\tau^2|\boldsymbol{t})$, the joint data model is

$$h(\boldsymbol{t}|\mu,\tau^2) = \left(\prod_{j=1}^{K}\frac{1}{[2\pi(\tau^2+V_j)]^{1/2}}\right)\exp\left[-\frac{1}{2}\sum_{j=1}^{K}\left(\frac{(t_j-\mu)^2}{\tau^2+V_j}\right)\right].$$

While we can work with either τ or τ^2, the induced prior for τ^2 is

$$\pi_3(\tau^2) = \frac{1}{2A(\tau^2)^{1/2}}\,I(0<\tau^2<A^2).$$

The full conditional posterior of τ^2 is then

$$\begin{aligned} p(\tau^2|\cdot) &\propto \pi_3(\tau^2)\,h(\boldsymbol{t}|\mu,\tau^2) \\ &\propto \frac{1}{(\tau^2)^{1/2}}\left(\prod_{j=1}^{K}\frac{1}{[2\pi(\tau^2+V_j)]^{1/2}}\right)\exp\left[-\frac{1}{2}\sum_{j=1}^{K}\left(\frac{(t_j-\mu)^2}{\tau^2+V_j}\right)\right]\,I(0<\tau^2<A^2). \end{aligned} \quad (3)$$

For use of the form $p(\mu, \tau^2, \boldsymbol{\theta}|\boldsymbol{t})$, the full conditional posterior of τ^2 is,

$$
\begin{aligned}
p(\tau^2|\cdot) &\propto \pi_3(\tau^2)\, g(\boldsymbol{\theta}|\mu,\tau^2) \\
&\propto \pi_3(\tau^2) \left[\prod_{j=1}^{K} [g(\theta_j|\mu,\tau^2)\right] \\
&\propto \frac{1}{(\tau^2)^{1/2}} \frac{1}{(\tau^2)^{K/2}} \exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{K}(\theta_j-\mu)^2\right] I(0<\tau^2<A^2) \\
&\propto \frac{1}{(\tau^2)^{(K+1)/2}} \exp\left[-\left(\frac{1}{2}\sum_{j=1}^{K}(\theta_j-\mu)^2\right)/\tau^2\right] I(0<\tau^2<A^2), \quad (4)
\end{aligned}
$$

which can be recognized as an Inverse Gamma density with parameters $(K-1)/2$ and $(1/2)\sum(\theta_j-\mu)^2$ truncated to have support $(0,\, A^2)$.

Question 4. The full conditional in (3) developed from the first form defies simplification and we would need a sub-algorithm such as a Metropolis-Hastings or some version of importance sampling to sample from $p(\tau^2|\cdot)$. On the other hand, a subalgorithm to sample from the density developed in (4) is farily simple, namely,

(a) At iteration $m+1$, given values $\mu^{(m)}$ and $\{\theta_j^{(m)} : j = 1,\ldots,K\}$, generate one value of a gamma random variable x with parameters $\alpha = (K-1)/2$ and $\beta = (1/2)\sum(\theta_j^{(m)} - \mu)^2$.

(b) Let $w = 1/x$.

(c) If $w < A^2$, let $(\tau^2)^{(m+1)} = w$, otherwise return to step 1 and sample a new x. Repeat until a value of w is accepted.

Question 5. A Wald interval for a binomial parameter p will be entirely on the positive line if

$$\hat{p} - 1.96\left(\frac{\hat{p}(1-\hat{p})}{n}\right)^{1/2} > 0.$$

With $\hat{p} = 3/n$ for binomial sample size n, this implies that

$$
\begin{aligned}
&\frac{3}{n} - 1.96\left[\left(\frac{3}{n}\right)\left(\frac{n-3}{n}\right)\left(\frac{1}{n}\right)\right]^{1/2} > 0 \\
\Rightarrow\ &\frac{9}{n^2} > 1.96^2\left(\frac{3n-9}{n^3}\right)
\end{aligned}
$$

$$
\begin{aligned}
&\Rightarrow \quad 9n - 3(1.96^2)n > -9(1.96^2) \\
&\Rightarrow \quad n < \frac{-9(1.96^2)}{9 - 3(1.96^2)} \\
&\Rightarrow \quad n < 13.694
\end{aligned}
$$

So for the lower endpoint to be on the positive line requires a binomial sample size of 13 *or less*, which is perhaps a bit surprising as we usually think that everything is improved as sample size increases. The key here, of course, is that with fixed $y = 3$, $\hat{p}$ decreases as n increases.

Note: Although not expected as part of the answer, the implicaiton of this result is that there is more to well-behaved Wald intervals than having the log likelihood surface well approximated by a quadratic near the maximum likelihood estimate (at least in the case of bounded parameter spaces). This latter (approxmation of the log likelihood by a quadratic) does improve as n increases, even with $y = 3$. But that is clearly not sufficient in this particular case to ensure that Wald intervals are contained in the parameter space.

Question 6. The quantities that were tested for normality in the Monte Carlo exercise reported in Table 1 of the question were

$$
\frac{\hat{p} - p}{[\hat{v}(\hat{p})]^{1/2}} = \frac{\hat{p} - p}{\left[\frac{\hat{p}(1-\hat{p})}{n}\right]^{1/2}} = \frac{n^{1/2}(\hat{p} - p)}{[\hat{p}(1-\hat{p})]^{1/2}}.
$$

These quantities are tested for a normal distribution with expected value 0 and variance 1.

Question 7. The issue is one of power in the Kolmogorov tests, which are based on sample sizes M. We are not provided information on how the $p-$values for the Kolmogorov tests were determined, so we don't know whether these are from simulation-based procedures or asymptotic approximation. We do know that the empirical distribution of the studentized values of $\hat{p}$ (see the answer to Question 6) coverges to the true distribution as M grows large. If the distribution is not dramatically different than standard normal, then small sample sizes may lead to the inability to distinguish the

Monte Carlo distribution from the theoretical normal, while this ability improves as the sample size increases.

Question 8. A normal prior for μ would likely be problematic unless we had a substantial amount of prior information at our disposal. Despite the fact that μ is still a location parameter, we need the posterior of mu to place negligible probability outside of the unit interval, and the same should be true of the prior predictive distribution of the θ_j. A normal prior having a variance we might consider as resulting in a prior that fails to be strongly informative (influential on the result) would violate this restriction. The obvious alternative is a beta prior (including, possibly, uniform).

Question 9. Formulating a prior on τ^2 that does not allow the prior predictive distribution of the θ_j to place more than negligible probability outside the unit interval may aslo prove to be a difficulty. Given that any beta random variable with possible values in the unit interval has variance less than 0.25, a uniform $(0, A)$ prior for τ^2 would be unreasonable if $A > 0.25$. If $\mu \neq 0.5$ this range becomes even smaller. So it is difficult to see how to assign a prior to τ^2 that does not either involve hard restrictions on support or allows too much probability on impossible regions in terms of the θ_j.

Question 10. The characteristic of estimators for model (1) that is lost with model (4) is that individual studies with smaller variability in the estimated effect (smaller V_j) have more influence on the point estimate of μ than do studies with greater variability (larger V_j). This is clearly the case for the maximum likelihood estimator of Question 1 and will also be true of the posterior mean for μ in a Bayesian analysis of model (1).

Question 11. Let the density of the V_j in model (5) be denoted as $p_j(v_j|\sigma^2)$ and the density of the T_j be denoted as $p(t_j|\theta_j, \sigma^2)$. The full conditional posterior of σ^2 in model (5) is then

$$p(\sigma^2|\cdot) \propto \pi(\sigma^2) \left[\prod_{j=1}^{K} p_j(v_j|\sigma^2)\right] \left[\prod_{j=1}^{K} p_j(t_j|\theta_j, \sigma^2)\right].$$

Here,

$$\begin{aligned}
\prod_{j=1}^{K} p_j(t_j|\theta_j,\sigma^2) &= \left[\prod_{j=1}^{K}\left(\frac{n_j}{2\pi\sigma^2}\right)^{1/2}\right] \exp\left[-\frac{1}{2}\sum_{j=1}^{K}\frac{n_j(t_j-\theta_j)^2}{\sigma^2}\right] \\
&= \frac{1}{(\sigma^2)^{1/2}}\left[\prod_{j=1}^{K}\left(\frac{n_j}{2\pi}\right)^{1/2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{K} n_j(t_j-\theta_j)^2\right].
\end{aligned}$$

Also,

$$\begin{aligned}
\prod_{j=1}^{K} p_j(v_j|\sigma^2) &= \left[\prod_{j=1}^{K}\left(\frac{n_j}{2\pi 2\sigma^2}\right)^{1/2}\right] \exp\left[-\frac{1}{2}\sum_{j=1}^{K}\frac{n_j(v_j-\sigma^2)^2}{\sigma^2}\right] \\
&= \frac{1}{(\sigma^2)^{1/2}}\left[\prod_{j=1}^{K}\left(\frac{n_j}{2\pi}\right)^{1/2}\right] \exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{K} n_j(v_j-\sigma^2)^2\right]
\end{aligned}$$

If $\pi(\sigma^2)$ is Inverse gamma with parameters α and β, then

$$\begin{aligned}
p(\sigma^2|\cdot) &\propto \frac{1}{(\sigma^2)^{\alpha+1}}\exp(-\beta/\sigma^2) \\
&\times \frac{1}{(\sigma^2)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{K} n_j(t_j-\theta_j)^2\right] \\
&\times \frac{1}{(\sigma^2)^{1/2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{K} n_j(v_j-\sigma^2)^2\right] \\
&\propto \frac{1}{(\sigma^2)^{\alpha+2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{j=1}^{K}\Big(n_j(t_j-\theta_j)^2+n_j(v_j-\sigma^2)^2\Big)\right],
\end{aligned}$$

which, due to the rightmost term in the exponential, cannot be coerced into the form of an inverse gamma density. So the full conditional posterior of σ^2 in model (5) is not amenable to the use of conditional conjugacy.

Question 12. The distinction between model (5) and model (6) is the amount of variation among the observed $V_j;\ j=1,\dots,K$. We could base a comparison of the models on the variability among these values in the actual data compares to values obtained by simulating from the posterior predictive distribution $p(\boldsymbol{V}^0|\boldsymbol{t},\boldsymbol{v})$ using the same sample sizes $n_j;\ j=1,\dots,K$ present in the actual data. For model (5) this is accomplished as, For each cycle $m = 1, \ldots, M$ in the MCMC algorithm used to sample from the joint posterior we have a value we can denote as $(\mu,\, \tau^2,\, \sigma^2,\, \{\theta_j : j = 1, \ldots, K\})_m$. For $j = 1, \ldots, K$,

(a) Sample $V^0_{j,m}$ from a normal distribution with expected value σ^2_m and variance $(2\sigma^2_m)/n_j$

Given a set of values $\{V^0_{j,m} : j = 1, \ldots, K\}$, compute the sample variance among these values as

$$S^2_m = \frac{1}{K} \sum_{j=1}^{K} (V^0_{j,m} - \bar{V^0}_m)^2,$$

where

$$\bar{V^0}_m = \frac{1}{K} \sum_{j=1}^{K} V^0_{j,m}.$$

If $S^{2,*}$ denotes the sample variance among the V_j in the actual data, compute

$$p = \sum_{m=1}^{M} I(S^2_m > S^{2,*}),$$

where $I(A)$ is the indicator function that assumes a value of 1 if A is true and 0 otherwise. We would then need to define some rule for interpretation of p. One option would be to rely on the concept of parsimony and take the position that the simpler model (5) should be preferred unless it can be demonstrated as inadequate to represent the variability among the observed V_j. Under this concept, we would state a preference for model (5) unless p was small (say less than 0.10 or 0.05).