Flourescence Resonance Energy Transfer (FRET) is a technique to gauge the interaction of two proteins at the molecular level. A FRET experiment involves affixing one protein to a surface and placing the second protein in a water solution. When the two proteins are interacting, a fluorescence molecule is activated and light of a particular wavelength is emitted. The experiment is run for 5 minutes and the emitted light is measured almost continuously. As a measure of interaction activity, scientist record the percentage of time the emitted light is above a threshold.

Part I. Pilot Study

Scientists run a pilot study to understand how adding an enzyme/activator into the water solution will affect activity. They create 8 samples and randomly assign 4 of those to receive the enzyme/activator.

control	7	26	0	27
enzyme	44	5	36	31

Table 1: FRET activity for samples with and without enzyme.

Scientists have decided to fit the effects model:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad \epsilon_i \overset{ind}{\sim} N(0, \sigma^2)$$

where, for observation i, Y_i is the observed FRET proportion and x_i is an indicator of the activator being present.

Using matrix notation, we have $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}^\top = (Y_1, Y_2, \ldots, Y_8)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top$.

1. Write a possible design matrix $\boldsymbol{X}$.

The scientists calculated the following quantities:

$$(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} = (15, 14)^\top \quad \text{and} \quad \boldsymbol{y}^\top (\boldsymbol{I} - \boldsymbol{X}[\boldsymbol{X}^\top \boldsymbol{X}]^{-1} \boldsymbol{X}^\top) \boldsymbol{y} = 1408$$

2. Calculate a t-statistic for a hypothesis test with null hypothesis that the mean FRET proportion is the same whether or not the activator is present, i.e. $H_0 : \beta_1 = 0$.

3. State the degrees of freedom for this test.

4. State whether the p-value will be less than or greater than 0.05. Explain why.

One scientist is worried about the equal variance assumption and suggests using Welch's two-sample t-test. A scientist scientist states that the results will be the same because the t-statistic will be the same.

5. Show that the second scientist is correct that the t-statistics will be the same whenever the sample sizes in the two groups are the same. For reference, the t-statistics for the two tests are

$$t_{\text{two-sample}} = \frac{\overline{y}_1 - \overline{y}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad \text{and} \quad t_{\text{Welch}} = \frac{\overline{y}_1 - \overline{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where, for group i with $i = 1, 2$, n_i is the number of observations, $\overline{y}_i$ is the sample mean, s_i^2 is the sample variance, and s_p is the pool standard deviation with

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

6. Explain why the results of the tests may be different even if the sample sizes in the two groups are the same.

This pilot study shows promise that the enzyme is increasing interaction activity. Thus the scientists would like to run a larger study to confirm their results.

7. Assuming a two-sample t-test will be used for the analysis with equal number of observations per group, calculate the number of observations per group to assure a type I error rate of 0.05 and a power of 0.8. The sample size formula is

$$n \geq \frac{2(z_{\alpha/2} + z_\beta)^2\sigma^2}{d^2}.$$

Part II. Meta-analysis

The scientist are interested in running a dose-response experiment where the amount of enzyme/activator is varied and the activity is measured. Before doing so, they perform a literature search to see what other scientists have measured. They find 6 manuscripts from 6 different labs. Each lab performed experiments that provide information about how enzyme amount affects activity. In these manuscripts, only the mean activity for each enzyme amount is reported. Table 2 provides the mean activity.

Table 2: Mean activity levels for 6 different labs with associated amount of enzyme.

	Enzyme amount									
lab	10	20	30	40	50	60	70	80	90	100
A	41	45	44	50	45					
B		22	24	39	38	27				
C			15	35	22	35	39			
D				16	23	24	31	15		
E					11	17	29	23	10	
F						0	17	13	27	16

8. Discuss the inferential statements that may be made from the results from these 6 labs including the concepts of generalization and causation.

The scientists first run a simple linear regression model and obtain the following estimates

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 39.4985233 5.55987000  7.104217 9.930614e-08
## amount      -0.2366758 0.09375599 -2.524381 1.753902e-02
```

9. Provide an interpretation for the coefficient for `amount`, i.e. the quantity -0.2366758.

They then run a multiple regression model including both amount and lab and obtain the following estimates

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  38.050572 4.30627645  8.836073 7.479014e-09
## amount        0.225432 0.09397066  2.398961 2.494115e-02
## labB        -17.004531 4.69853281 -3.619115 1.440936e-03
## labC        -19.992615 4.97245974 -4.020669 5.343908e-04
## labD        -29.785971 5.39820322 -5.517757 1.303502e-05
## labE        -35.629936 5.94322614 -5.995050 4.105282e-06
## labF        -41.394802 6.57794594 -6.292968 2.019958e-06
```

10. Provide an interpretation for the coefficient for `amount`, i.e. the quantity 0.225432.

11. Explain how the coefficient for `amount` can be negative and significant (at the 0.05 level) in the simple linear regression model yet positive and significant (at the 0.05 level) in the multiple regression model.

Both the simple linear regression and the multiple regression models assume **equal variance**.

12. Describe why this assumption may be incorrect.

13. What additional summary statistics (other than the data itself) could have been provided in the manuscripts that would allow you to address this assumption?

Part III. Dose-response

The scientists ultimately run their own dose-response experiment where they randomly assigned the enzyme amount and measured the interaction activity. Table 3 contains summary statistics

Enzyme Amount	N	Mean	SD
0	5	14	8
50	5	24	3
100	5	37	9
150	5	82	9
200	5	76	9

Table 3: Dose-response experiment: Activity summary statistics.

for each enzyme amount including sample size (N), sample mean of activity (Mean), and sample standard deviation of activity (SD).

The scientists first assess the scientific question 'How does enzyme presence (versus absence) affect interaction activity?'

14. Let $\mu_0, \mu_{50}, \mu_{100}, \mu_{150}$, and μ_{200} be the true mean activity for enzyme amounts 0, 50, 100, 150, and 200, respectively. State a contrast in terms of these means that would answer the scientific question.

15. Construct a 95% confidence interval for this contrast.

Table 4 provides residual sums of squares for 4 models: intercept-only, enzyme amount treated as continuous, enzyme amount treated as continuous but including a quadratic term, and enzyme amount treated as a categorical variable.

Model	Residual Sums of Squares
Intercept-only	20081
Amount (continuous)	3558
Amount (quadratic)	3557
Amount (categorical)	1203

Table 4: Dose-response experiment: Residual sums of squares for another 4 different models.

16. In the quadratic model, will the p-value for the coefficient for the quadratic term be large or small? Explain your answer.

17. Calculate the F-statistic for the lack-of-fit test for the model where enzyme amount is treated as a continuous variable. State the numerator and denominator degrees of freedom. Will the p-value for this test be large or small? Explain your answer.

Part IV. Piecewise Mean

A colleague suggests to the scientists that they fit the following model

$$Y_i = \mu(x) + \epsilon_i,\ \epsilon_i \overset{ind}{\sim} N(0, \sigma^2), \quad \mu(x) = \begin{cases} \beta_0 + \beta_1 x & x < \tau \\ \gamma & x \geq \tau \end{cases}$$

This model has a piecewise function for the mean consisting of a linear function up to τ and a constant function after τ.

18. What restriction is necessary so that the function $\mu(x)$ is continuous?

19. If τ is known, explain how you would fit this function using a multiple regression model.

20. If τ is unknown, explain how you would fit this model including how you would obtain uncertainty about τ.

Part I: Pilot Study

1. One possible design matrix is

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Other possible matrices are found by rearranging the rows.

2. The t-statistic for this test is

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}} = \frac{14 - 0}{\sqrt{1408/(8-2)}} = 0.9139077$$

3. The degrees of freedom is 8 - 2 = 6.

4. The p-value for this test will be larger than 0.05 since this statistic is much less than 1.96 (the cutoff for significance of a standard normal).

5. The test statistic for the test of $H_0 : \beta_1 = 0$ is equivalent to the two-sample t-test. So the test that has been performed is equivalent to the two-sample t-test.

Since the numerators are the same in the two-sample and Welch t-tests, it suffices to show the denominators are the same when $n_1 = n_2 = n$.

$$s_p^2 = \frac{(n-1)s_1^2 + (n-1)s_2^2}{n+n-2} = \frac{(n-1)s_1^2 + (n-1)s_2^2}{2(n-1)} = \frac{s_1^2 + s_2^2}{2}$$

The two-sample t-statistic denominator is then

$$s_p\sqrt{\frac{1}{n} + \frac{1}{n}} = \sqrt{\frac{s_p^2}{n} + \frac{s_p^2}{n}} = \sqrt{\frac{s_1^2 + s_2^2}{2n} + \frac{s_1^2 + s_2^2}{2n}} = \sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}$$

which is the denominator of Welch's t-statistic when $n_1 = n_2 = n$.

6. The p-values (and therefore the results and decisions) may be different because the degree of freedom in the two-sample t-test is $n_1 + n_2 - 2$ while the Welch's t-test uses the Satterthwaite degrees of freedom which generally won't be equal to $n_1 + n_2 - 2$.

7. The sample size formula is

$$n = \frac{2(z_{\alpha/2} + z_\beta)^2\sigma^2}{\delta^2} = \frac{2(1.96 + 0.84)^2 \times 15.3^2}{14^2} = 18.7272$$

Round up to 19 for the sample size required per group.

Part II: Meta-analysis

8. There are two main concerns with generalizability: inference to a greater population and causal inference.

It is unlikely that these results are able to be generalized to any larger population as the labs (and their materials) are unlikely to be a random sample from any population. In addition, the labs seem to have some distinct differences since, for a given enzyme amount, the mean activity level is almost always decreasing.

For causal inference we are concerned with making causal claims about the enzyme amount affecting activity. It is not stated, but it is possible the labs randomly assigned enzyme amount and therefore can make claims about enzyme amount **causing** changes in mean activity. If random assignment of enzyme amount was performed, the causal claim would still be limited within the lab.

9. For each unit increase in amount, the mean activity is expected to decrease by 0.2366758.

10. While holding lab constant, for each unit increase in amount, the mean activity is expected to increase by 0.225432.

11. This is an example of Simpson's Paradox. When we ignore lab, there appears to be a decrease in mean activity when we increase enzyme amount. This can be seen in Table 2 by taking an average within each enzyme amount. When we control for lab, we see (on average) increases in activity as enzyme amount increases.

The key is that there is a relationship between lab and enzyme amount. Labs indicated with letters later in the alphabet use larger amounts of enzyme but simultaneously have less activity. Perhaps there are systematic differences amongst the labs that is causing this to happen, e.g. different measuring apparatus or measuring interaction of different proteins.

12. The uncertainty in the mean activity may vary quite a bit amongst the different combinations of lab and enzyme amount.

13. If the labs had provided a standard error (or standard deviation with sample size) for each of the values in Table 2, then we could have assessed the equal variance assumption. In particular, we could have performed a weighted regression analysis.

Part III: Dose-response

14. An appropriate contrast is

$$\gamma = \mu_0 - \frac{\mu_{50} + \mu_{100} + \mu_{150} + \mu_{200}}{4}.$$

15. The point estimate is

$$\hat{\gamma} = 14 - \frac{24 + 37 + 82 + 76}{4} = -40.75.$$

The standard error is

$$SE(\hat{\gamma}) = \hat{\sigma}\sqrt{\left(\frac{1}{5} + 4 \times \frac{(-1/4)^2}{5}\right)}.$$

The error variance estimate can be found using the enzyme amount standard deviations in Table 3 but an easier approach is to use the last row of Table 4:

$$\hat{\sigma} = \sqrt{1203/(25-5)} = 7.7556431.$$

Thus, our standard error is

$$SE(\hat{\gamma}) = \hat{\sigma}\sqrt{\left(\frac{1}{5} + 4 \times \frac{(-1/4)^2}{5}\right)} = 3.8778216.$$

Finally, we have a t-critical value with 20 degrees of freedom at $\alpha = 0.05$ which is ≈2.09. Thus our 95% CI for this contrast is (-48.838994, -32.661006).

16. The residual sums of squares only dropped 1 when adding this term to the model indicating the quadratic term did not improve model fit much. Thus, the p-value will be large, i.e. close to 1.

17. The F-statistic is

$$F = \frac{(3558 - 1203)/(4-1)}{\hat{\sigma}} = 101.2166224.$$

This F-statistic has 3 numerator and 20 denominator degrees of freedom. A guideline for significance (at the 0.05 level) of F-statistics is around 4. As this value is much greater than 4, the p-value will be small indicate that the model where amount is treated as continuous is not a very good fit to the data.

Part IV: Dose-response

18. For the function to be continuous, we need $\gamma = \beta_0 + \beta_1\tau$.

19. If τ is known, we can set all enzyme amounts (x) above τ to be equal to τ. Then we run a multiple regression analysis using these modified enzyme amounts.

20. One approach would be to run a parametric bootstrap using the maximum likelihood estimator for all parameters. Another approach would be to perform a Bayesian analysis, e.g. a Gibbs sampler that alternates drawing τ from its full conditional and the remaining parameters from their full conditional.

Part I

The data for this problem are from a study of baking chocolate cakes. The researcher evaluated two factors: recipe (A, B, or C) and baking temperature (175, 185, 195, 205, 215 and 225 degrees Celsius) in all combinations. She made a batch of cake batter using one of the 3 recipes then divided it into 6 cake tins. The 6 possible baking temperatures were randomly assigned to the 6 cake tins. The experiment was conducted over 15 days. Three batches of batter, one of each recipe, were made and baked each day. Recipe can be considered randomly assigned to batch. The response we consider is a measured value called the breaking angle. Details of the measurement are not important.

There are 15 days, 3 recipes, 6 oven temperatures, 45 batches of batter and 270 observations. There are no missing data.

One possible model is:

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + r_j + t_k + rt_{jk} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma_e^2). \end{aligned} \tag{1}$$

Y_{ijk} is the measured angle for the cake made with recipe j and baked on day i at temperature k,
α_i is the main effect of day i,
r_j is the main effect of recipe j,
t_k is the main effect of temperature k, and
rt_{jk} is the interaction effect of recipe j and temperature k.

1. Complete the skeleton ANOVA table with appropriate names for sources of variation and their degrees of freedom (df).

Source	df
Days	14
⋮	⋮

You adopt a non-full rank parameterization for model (1). A computer software program reports these estimates for some of the parameters and some of the combinations of parameters:

Parameter	μ	α_1	$\Sigma_{i=1}^{15}\alpha_i$	r_1	r_2	r_3	$\Sigma_{j=1}^{3}r_j$	t_1	t_2
Estimate	29.44	19.5	4.69	-0.16	0.04	0.50	0.38	-7.80	-6.80

Parameter	rt_{11}	rt_{12}	rt_{21}	rt_{22}	rt_{31}	rt_{32}
Estimate	1.87	3.27	-0.60	0.93	0.00	0.00

2. Are these estimates unique? Briefly explain why or why not.

3. One quantity of potential interest is $\tau = \mu + \Sigma_{i=1}^{15}\alpha_i/15 + (r_1 + r_2 + r_3)/3 + t_1 + (rt_{11} + rt_{21} + rt_{31})/3$. In terms of the factors used in this study, describe what τ represents.

4. Is τ estimable? Briefly explain why or why not.

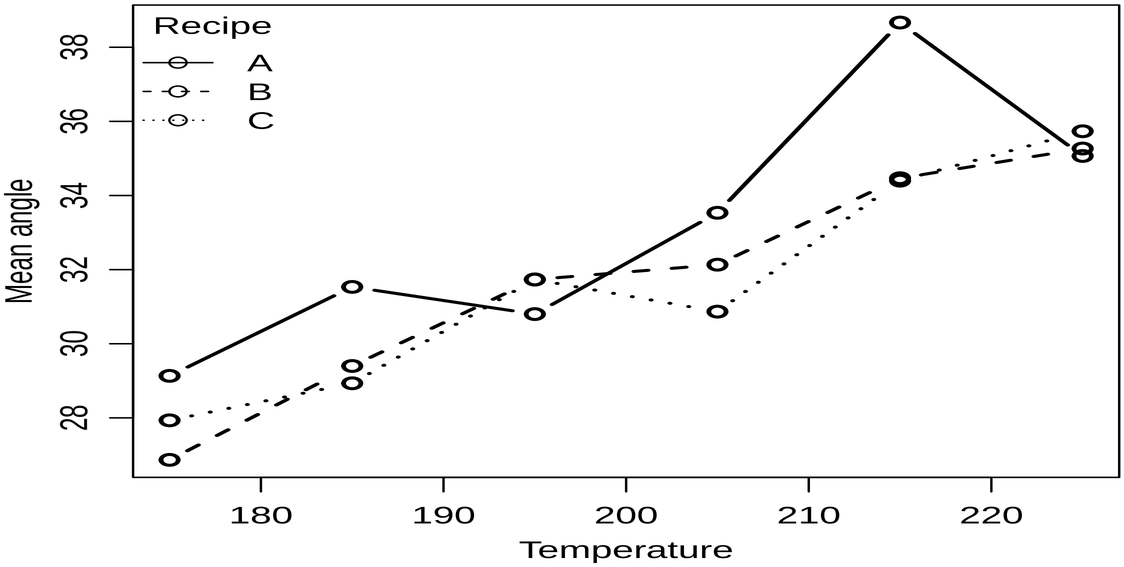
5. Estimate the mean breaking angle for recipe A ($j = 1$) baked at 185 degrees ($k = 2$). If additional information is needed, describe what is needed.

6. Your computer software program provides Type I (sequential) Sums-of-Squares (SS) and F statistics. This information for Temperature is:

Source	Type I SS	F
Temperature	2100.3	18.186

   You want type III (partial) statistics. Report the type III F statistic for Temperature, if possible given the information provided. If not, what additional information is needed?

7. The plot below shows angle means for each combination of temperature and recipe.



   This plot suggests a linear response to temperature with the same slope for each recipe. One model to evaluate this is

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + r_j + \beta T_{ijk} + \gamma_j T_{ijk} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma_e^2). \end{aligned} \tag{2}$$

   T_{ijk} is the temperature used for the cake made with recipe j and baked on day i at temperature k. $\beta + \Sigma_{j=1}^{3}\gamma_j/3$ is the temperature slope averaged over the three recipes.

   A second approach is to fit model (1) and estimate the marginal mean for each temperature, $\overline{Y}_{..k}$. Then, evaluate the linear trend contrast, $\Sigma_{k=1}^{6} c_k \overline{Y}_{..k}$ with coefficients, c_k = -5, -3, -1, 1, 3, 5 for temperatures in ascending order.

   Are these two approaches equivalent, i.e., will they give identical results for the linear temperature trend in breaking angle, averaged over recipes? If not, describe what quantities in the analysis will differ.

8. You want to test for lack of fit of the temperature part of model (2), i.e., whether a linear trend adequately describes the relationship between temperature and breaking angle. Some additional information from models (1) and (2) that may be useful:

Quantity	Model	Estimate
Linear slope, $\hat{\beta}$	(2)	0.158
se of $\hat{\beta}$	(2)	0.0172
MSE	(2)	23.34
SS for linear trend	(1)	1966.7
MSE	(1)	23.10

Test the null hypothesis of no lack of fit for a linear trend/linear regression model. Report the test statistic and its distribution under the null hypothesis. A p-value is not needed. If not possible with the provided information, describe briefly what additional information you need.

9. If the p-value for the test in question **8**. is 0.35, is it appropriate to conclude that the linear regression fits the data? If not, provide a more appropriate conclusion.

A colleague points out that 6 cakes were made from the same batch of cake batter. Note that the combination of day, i, and recipe, j, uniquely identifies each batch of cake batter.

10. Does this information suggest any changes to model (1)? If so, write a more appropriate model. Please use subscripts i for day, j for recipe, and k for temperature. Include distributions for random effects.

11. Explain why your model in question **10**. is more appropriate OR why no changes are needed.

Your colleague then points out that the same 6 ovens were used throughout the study and suggests there are consistent differences among ovens. E.g., oven 1 consistently behaves differently from the other ovens, no matter what temperature it is set to or what type of batter is baked in it. Based on this, you decide to include oven effects in the model.

12. The investigators were very careful to re-randomize temperatures to ovens for each batch of batter, so, e.g., they avoided always using oven 1 at 185 degrees. Briefly explain why re-randomization was essential.

13. If you add an oven effect to the model, is the oven effect nested within the batter effect, or crossed with the batter effect? Briefly explain your answer.

14. If want to make inferences about recipe and temperature effects for these 6 ovens, should differences among ovens be considered a fixed effect or a random effect? Briefly explain your answer.

Part II

These questions explore the fate of whaling ships: did the ship return to port or was it missing? The data are information on 5,000 whaling trips randomly sampled from a much larger data set for all recorded whaling trips from US ports between 1688 and 1937. There is one row of data for each departure from a US port.

The response variable is missing. ‘No’ indicates that the ship returned to port on that trip; ‘Yes’ indicates the ship was missing.

The captain’s wife accompanied the captain on some trips. The variable wife has the values: ‘No’, the wife was not on that trip, or ‘Yes’, the wife was. Our questions consider different approaches to explore the association between wife and missing.

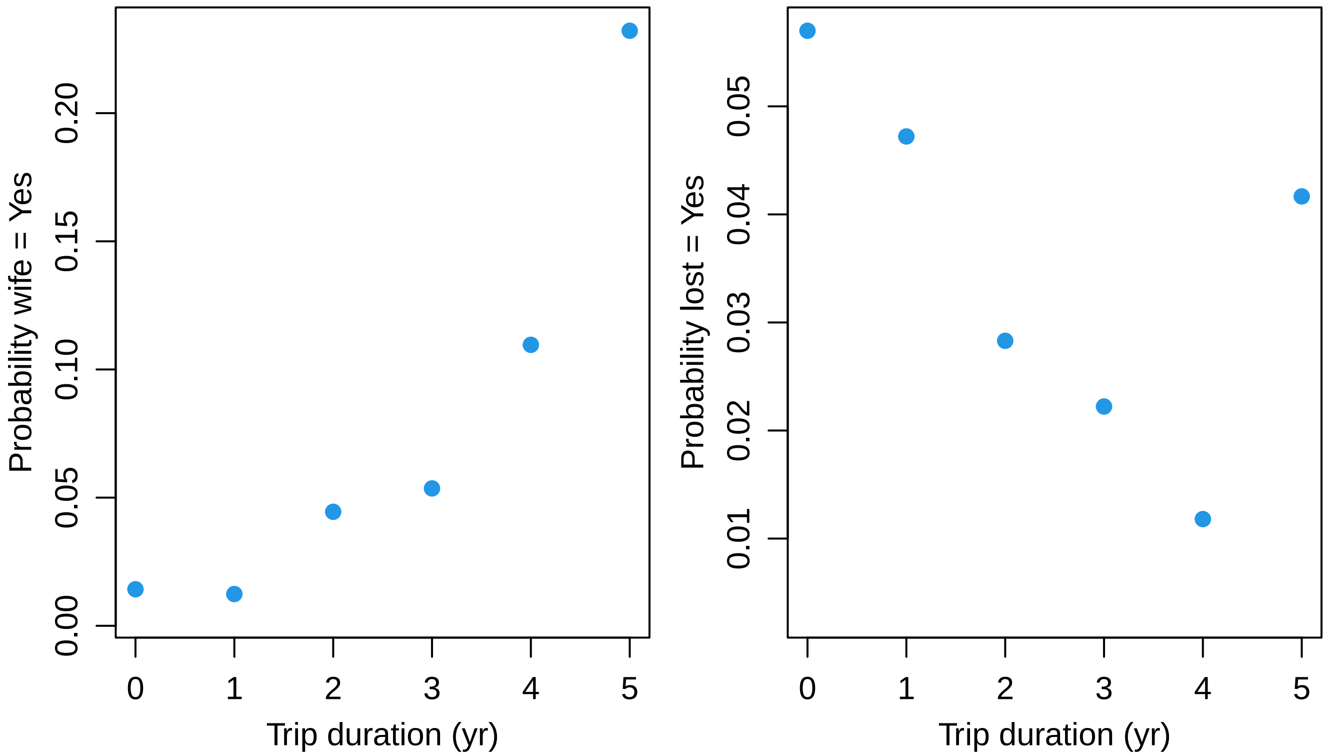
The first approach is based on the 2 x 2 contingency table:

	missing		
wife	No	Yes	Total
No	4649	143	4792
Yes	192	16	208
Total	4841	159	

For **Questions 15-17**, you can assume that each trip is an independent observation.

15. Estimate the odds ratio and its 95% confidence interval. Express the odds ratio as (odds of missing when wife present) / (odds of missing when wife not present).

Whaling trips during this era could last 5 years or more. Because of small sample sizes, durations were rounded to the closest year and all trips longer than 5 years were included with those of 5 years. These two plots suggest trip duration is associated with both the probability that wife = ‘Yes’ and the probability that the ship was missing.



Note: There are relatively few trips of 5 years or more.

16. You fit a logistic regression model using a spline function of duration to predict the probability of missing = 'Yes'. Indicator variable *missing* is 1 when missing = 'Yes' and 0 when missing = 'No'. The R code is:

```
trips.m1 <- gam(missing ~ s(duration, k=4),
  data=trips,
  family=binomial)
```

Note: the $k = 4$ option sets the number of knots in the spline.

The estimated degrees of freedom (edf) for the spline term is 2.733. What, if anything, does this tell you about the form of the relationship between duration and the logit transformed probability of missing? In particular, does it suggest anything about the appropriateness of a linear relationship, i.e., logit $P[\text{missing}] = \beta_0 + \beta_1\text{duration}$?

17. You add the indicator variable *wife1* to the model in **Question 16**. It has the value of 1 when wife = 'Yes' and 0 when wife = 'No'. The estimated regression coefficient for *wife1* = 1.267 with a standard error of 0.297. Estimate the associated odds ratio and its 95% confidence interval. Express the odds ratio as (odds of missing when wife present) / (odds of missing when wife not present).

The previous questions in **Part II** assume that each trip is an independent observation. However, a single vessel and captain can make multiple trips, which should be expected to introduce correlations among observations for that vessel. The median number of trips per vessel is 2, but a few vessels make over 20 trips.

18. One frequently used approach to account for such correlations is to add a random effect for vessel, i.e.,

$$\begin{aligned} L_{ij} \mid \tau_i &\overset{independent}{\sim} \text{Bernoulli}\,(\pi_{ij}) \text{ with logit}\,(\pi_{ij}) = \beta_0 + \beta_1 W_{ij} + f(D_{ij}) + \tau_i \\ \tau_i &\overset{iid}{\sim} N(0, \sigma^2_\tau), \end{aligned} \qquad (3)$$

i indexes vessels, j indexes trips within vessels, L_{ij} is the indicator variable for whether the ship went missing, W_{ij} is the indicator variable for whether the captain's wife was present, $f(D_{ij})$ is the spline function for duration of the trip, and τ_i is a vessel-specific shift in the intercept. Briefly explain why the log likelihood for model (3) is difficult to compute.

A different approach to account for multiple observations from the same vessel is to condition on the total number of events (missing) for each vessel. Define the total number of times the vessel i went missing as $L_i^* = \Sigma_j L_{ij}$. Then, write the likelihood conditional on L_i^*. For these data, L_i^* has one of three values, 0, 1, or 2. (Yes, in these data, a vessel could be declared missing on more than 1 trip. That was rare).

The remaining questions explore this conditional approach. For simplicity, we will ignore vessels with only one trip, consider only the first two trips for each remaining vessel, and ignore duration. The model for vessel i on trip j, $j = 1, 2$ is now:

$$\begin{aligned} L_{ij} \mid \tau_i &\overset{independent}{\sim} \text{Bernoulli}\,(\pi_{ij}) \text{ with logit}\,(\pi_{ij}) = \beta_0 + \beta_1 W_{ij} + \tau_i \\ \tau_i & \quad \text{unspecified.} \end{aligned} \qquad (4)$$

We assume that $L_{ij} \mid \tau_i$ are conditionally independent.

For the next few questions, use the following notation in your answers:
$\boldsymbol{X}_{ij}$ is a row vector with the covariate values for trip j of vessel i, i.e., $\boldsymbol{X}_{ij} = [1,\ W_{ij},\ 1]$
$\boldsymbol{\beta}_i$ is the coefficient vector for vessel i, i.e. $\boldsymbol{\beta}_i = [\beta_0,\ \beta_1,\ \tau_i]$
Hence, equation (4) can be written as logit $\pi_{ij} = \boldsymbol{X}_{ij}\boldsymbol{\beta}_i$
Define $[L_{ij} \mid \tau_i,\ X_{ij}]$ as the distribution of L_{ij} conditional on τ_i and $\boldsymbol{X}_{ij}$
To further simplify notation, the conditioning on $\boldsymbol{X}_{ij}$ will be omitted henceforth.

19. State the joint conditional pmf $[L_{i1},\ L_{i2} \mid \tau_i]$ in terms of the π_{ij} used in equation (4)

20. Derive the conditional pmf $[L_{i1} = 0,\ L_{i2} = 0 \mid L_i^* = 0,\ \tau_i]$

21. Derive the conditional pmf $[L_{i1} = 0$ and $L_{i2} = 1 \mid L_i^* = 1,\ \tau_i]$ and express it in terms of $\beta_0,\ \beta_1,$ and W_{ij}.

22. The third conditional pmf $[L_{i1} = 1,\ L_{i2} = 1 \mid L_i^* = 2,\ \tau_i]$ has the same form as that derived in Question **20**. Based on the expressions in **Questions 20**. and **21**., explain why working with the conditional likelihood avoids the computational issues associated with model (3).

Part 1. Baking cakes

1.

Source	df
Days	14
Recipe	2
Temperature	5
R*T	10
Residual	238

2. No. The parameterization is not full rank, so there are an infinite number of solutions.

3. τ is the mean angle for temperature 1 (175 degrees), averaged over recipes and replicates. Note: would be a good idea to include the phrase "marginal mean"

4. Yes, it is estimated by the average of the 45 observations in all 15 replicates of recipes A, B, and C. This is an average of the observations, so it is estimable.

5. 26.06. Computed as $\mu + \Sigma_{i=1}^{15} \alpha_i / 15 + r_1 + t_2 + rt_{12} = 29.44 + 4.69/15 - 0.16 - 6.8 + 3.27$.

6. The type III F is the same as the type I F, 18.186. The data are balanced, so type I and type III SS are the same, as are the type I and type III F statistics.

7. Some results are identical: SS for the regression slope = SS for the linear contrast.
Some results are proportional: estimated regression slope and estimated value of the linear contrast
Others are not, at least almost certainly: estimated error variance. T or F statistics and p-values because they depend on the error variance

8. SS for lack of fit = SS for temperature groups - SS for linear trend. The question was unclear whether lack of fit was just for the temperature term or for both the temperature term and the temperature*recipe interaction. The first uses the SS associated with just temperature; the second uses the error SS.
Using SS for temperature: Question 6 gives you SS for temperature groups = 2100.3; this question gives you SS for linear trend = 1966.7. So SS for lack of fit = 2100.3 - 1966.7 = 133.6, with (5-1) = 4 df. F = (133.66 / 4) / 23.34 = 1.43. Has a central F distribution with 4, 238 df.
Using error SS: This question gives you MSE values for temperature groups (model 1) and linear trend (model 2). The error has 238 df for the temperature groups model and 250 df for the linear trend model so the error SS values are 23.10*238 = 5497.8 for temperature groups and 23.34*250 = 5835.0 for the linear trend models. So this SS for lack of fit = 5835.0 - 5497.8 = 133.6 with (250 - 238) = 12 df. F = (133.6/12) / 23.34 = 0.48. Has a central F distribution with 12, 238 df.
Note: Testing whether the temperature slope = 0 is not a test of lack of fit.

9. No. "No evidence of lack of fit for a linear trend". Other wordings possible, including "lack of fit is not statistically significant".

10. Yes,

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + r_j + \tau_{ij} + t_k + rt_{jk} + \varepsilon_{ijk}, \qquad (1) \\ \tau_{ij} &\overset{iid}{\sim} N(0, \sigma_b^2). \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma_e^2). \end{aligned}$$

Note: the combination of i (replicate) and j (recipe) identifies each batch of batter.

11. An appropriate model needs to include variability between batches. (may have been included in the answer to the previous question).

12. To avoid confounding temperature effects with oven effects

13. An effect specific to each oven is crossed with batter. The same oven is used with all batches and the problem text tells you that the effect of that oven might be consistently different from the effect of a different oven. That means that if there is any difference associated with an oven, it is consistent across all batches.
Note: a nested effect for ovens would have a different effect for each batter.

14. Ovens should be a fixed effect because inferences will be about these 6 ovens.
Note: If ovens were a random effect, the analyses would treat the 6 ovens in the study as a simple random sample from a (almost certainly hypothetical) population of ovens.

Part 2 Whaling ships.

15. 2.71, with 95% ci of (1.58, 4.62)
Details of the computations:
Odds of missing with wife = 16/192 = 0.0833, without wife = 0.0308, so odds ratio = 0.0833 / 0.0308 = 2.71.
Standard error of log odds = $\sqrt{1/4649 + 1/143 + 1/192 + 1/16} = \sqrt{0.0749} = 0.274$.
95% ci for log odds = $\log 2.71 \pm 1.96 \times 0.274 = 0.997 \pm 0.536 = (0.46, 1.53)$.
So 95% ci for odds ratio = $(\exp 0.46, \exp 1.53) = (1.58, 4.62)$

16. It suggests that the relationship is more flexible than a straight line (edf = 1). The relationship is closer to a cubic (edf = 3) than quadratic (edf = 2).

17. Odds ratio = $\exp(\beta) = \exp(1.267) = 3.55$.
95% ci for the regression coefficient is $1.267 \pm 1.96 \times 0.297 = 1.267 \pm 0.582 = (0.68,\ 1.85)$.
Hence, the 95% ci for the odds ratio is $\exp(0.68,\ 1.85) = (1.98,\ 6.35)$.

18. The log likelihood requires the marginal distribution of L_{ij} for specified values of the parameters ($\beta_0,\ \beta_1,\ f(D_{ij})$, and σ_τ^2). This marginal likelihood requires integrating over the distribution of τ_i. There is no analytic expression for that integral, so it must be approximated, e.g. by a Laplace approximation or by linearizing the log likelihood expression.

19. Conditional on τ_i, L_{i1} and L_{i2} are independent, so their joint conditional probability = $\pi_{i1}^{L_{i1}}(1-\pi_{i1})^{(1-L_{i1})}\pi_{i2}^{L_{i2}}(1-\pi_{i2})^{(1-L_{i2})}$

20. When $L_i^* = 0$, $L_{i1} = L_{i2} = 0$, so
$[L_{i1} = 0,\ L_{i2} = 0 \mid \tau_i, L_i^* = 0] = \frac{[L_{i1}=0,\, L_{i2}=0,\, L_i^*=0|\tau_i]}{[L_i^*=0|\tau_i]} = \frac{[L_{i1}=0,\, L_{i2}=0|\tau_i]}{[L_{i1}=0,\, L_{i2}=0|\tau_i]} = 1.$

21. We start by deriving $P = [L_{i1} = 0,\ L_{i2} = 1,\ L_i^* = 1 \mid \tau_i,\ L_i^* = 1]$, then express that probability in terms of the parameters of the logistic regression model.

$$\begin{aligned}
P &= [L_{i1} = 0, L_{i2} = 1,\ L_i^* = 1 \mid \tau_i,\ L_i^* = 1] \\
&= [L_{i1} = 0 \mid t_i][L_{i2} = 1 \mid \tau_i]/[L_i^* = 1 \mid \tau_i] \\
&= \frac{[L_{i1} = 0 \mid \tau_i][L_{i2} = 1 \mid \tau_i]}{[L_{i1} = 0 \mid \tau_i][L_{i2} = 1 \mid \tau_i] + [L_{i1} = 1 \mid \tau_i][L_{i2} = 0 \mid \tau_i]} \\
&= \frac{1}{1 + [L_{i1} = 0 \mid \tau_i][L_{i2} = 1 \mid \tau_i]/[L_{i1} = 0 \mid t_i][L_{i2} = 1 \mid \tau_i]} \\
&= 1/(1 + R) \\
\text{where } R &= [L_{i1} = 0 \mid \tau_i][L_{i2} = 1 \mid \tau_i]/[L_{i1} = 0 \mid \tau_i][L_{i2} = 1 \mid \tau_i]
\end{aligned}$$

$$\begin{aligned}
R &= \frac{\dfrac{1}{1 + \exp(\boldsymbol{X}_{i1}\boldsymbol{\beta})} \dfrac{\exp(\boldsymbol{X}_{i2}\boldsymbol{\beta}_i)}{1 + \exp(\boldsymbol{X}_{i1}\boldsymbol{\beta}_i)}}{\dfrac{\exp(\boldsymbol{X}_{i1})}{1 + \exp(\boldsymbol{X}_{i1}\boldsymbol{\beta})} \dfrac{1}{1 + \exp(\boldsymbol{X}_{i2})\boldsymbol{\beta}_i)}} \\
&= \frac{\exp(\boldsymbol{X}_{i2}\boldsymbol{\beta}_i)}{\exp(\boldsymbol{X}_{i1}\boldsymbol{\beta}_i)} \\
&= \frac{\exp(\beta_0 + \beta_1 W_{i2} + t_i)}{\exp(\beta_0 + \beta_1 W_{i1} + t_i)} \\
&= \frac{\exp(\beta_0 + \beta_1 W_{i2})}{\exp(\beta_0 + \beta_1 W_{i1})} \\
\text{since } P &= 1/(1 + R) \\
P &= \frac{\exp(\beta_0 + \beta_1 W_{i1})}{\exp(\beta_0 + \beta_1 W_{i1}) + \exp(\beta_0 + \beta_1 W_{i2})}
\end{aligned}$$

22. None of the conditional probabilities, $[L_{i1}, L_{i2} \mid L_i^* = 0]$, $[L_{i1}, L_{i2} \mid L_i^* = 1]$ or $[L_{i1}, L_{i2} \mid L_i^* = 2]$ involves the unknown t_i. The conditional likelihood can be maximized without needing to integrate out the distribution of t_i or optimizing over many nuisance parameters (all the t_i).
Note: using a conditional likelihood requires considering only the subset of data from the vessels that were missing. Conceptually, the conditional likelihood compares the probability that a wife was on board between the trips prior to missing and the trip that had the missing. You may recognize this result as the probability used in McNemar's test. The conditional likelihood approach generalizes to more than 2 observations in a match group.

Limnology is the study of chemical, biological, and physical features of lakes and other bodies of freshwater. Limnologists are interested in the relations between phosphorus (one of the primary plant nutrients) and the level of algal growth (the main source of primary production) in these waters. We will be concerned here with one particular study in which 15 lakes and reservoirs in the Midwestern United States were each sampled once a week between June 1 and October 31. There are a total of $n = 330$ observations in the available data and the quantities we will be concerned with are the concentration of total phosphorus (in mg/L) and the concentration of chlorophyll (in μg/L). Chlorophyll is a measure of algal growth in lakes.

The typical statistical analysis used to relate concentrations of chlorophyll and phosphorus is regression of chlorophyll as a response on phosphorus as a covariate. Such chlorophyll-phosphorus regressions may use various scales for the variables, the most common being logarithmic transformations of both chlorphyll and phosphorus. Also, in many limnological studies, lakes in a region are sampled a number of times between late spring and early fall. Such samples are then used to compute *seasonal mean* values of the measured variables. In regressions to relate chlorophyll to phosphorus using seasonal mean data, each point then represents one lake. Thus, there are two data manipulations that are often used, (1) transformation of measured variables and (2) aggregation of data values. The data we are concerned with here were collected as part of a study to examine the effects of these data manipulations on regression analysis.

To examine the issue of data aggregation, we will consider three potential regression models for the logarithm of chlorophyll versus the logarithm of phosphorus. Consider our original observations to consist of multiple data values for each of a number of lakes, $\{Y_{g,i} : \; g = 1, \ldots, G; \; i = 1, \ldots, n_g\}$, where g indexes lake and i indexes observation within lake. The three models we want to consider at this point are (1) a simple linear regression model for all individual observations, $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$,

$$\log(Y_{g,i}) = \beta_0 + \beta_1 \log(x_{g,i}) + \sigma\epsilon_{g,i}, \tag{1}$$

a simple linear regression model for seasonal mean values $W_g = (1/n_g)\sum_i \log(Y_{g,i})$ versus $z_g = (1/n_g)\sum_i \log(x_{g,i})$, $g = 1, \ldots, G$,

$$W_g = \alpha_0 + \alpha_1 z_g + \tau\epsilon_g \tag{2}$$

and the collection of simple linear regression models for observations within each lake, for $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$,

$$\log(Y_{g,i}) = \beta_{g,0} + \beta_{g,1} \log(x_{g,i}) + \sigma_g\epsilon_{g,i}. \tag{3}$$

Note that the additive error terms $\epsilon_{g,i}$ in (1), ϵ_g in (2) and $\epsilon_{g,i}$ in (3) may be distinct and the notation is is not intended to imply that they are necessarily related to one another. Also note that the study design was to sample each lake once per week, which is known to be sufficient to render an assumption of independence among observations within a lake a reasonable assumption.

A scatterplot of log chlorophyll versus log total phosphorus for the original data values from our 15 lakes is shown in Figure 1, along with the fitted expectation function from the simple linear regression model (1) fitted to these data. A scatterplot of seasonal mean log chlorophyll versus seasonal mean log phosphorus for these same lakes is shown in Figure 2, along with the estimated expectation function from model (2). Estimates corresponding to these fitted regressions are given in Table 1.

Data Form	n	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	r^2
Original Obs. (Fig 1)	330	-1.968	1.273	0.342	0.732
Seasonal Means (Fig 2)	15	-1.863	1.280	0.109	0.901

Table 1: Summaries of regressions shown in Figures 1 and 2.

1. The estimated regression coefficients in Table 1 are quite similar for regressions with the two forms of data, which correspond to models (1) and (2). Give a sufficient condition on model (3) that will ensure regression coefficients β_0 and β_1 in model (1) will be equal to α_0 and α_1 in model (2). You may assume that all of the $\epsilon_{g,i}$ and $x_{g,i}$ in (1) and (2) are mutually independent.
   Note: this question asks for a sufficient condition only, not a necessary condition.

2. Regardless of whether the regression coefficients are the same or not,

   a. If both (3) and (2) hold, what must be true of the σ_g^2, the n_g, and τ^2?

   b. If both (3) and (1) hold, what must be true of the σ_g^2, the n_g, and σ^2?

   c. If (2) and (1) hold, what must be true of the n_g, τ^2 and σ^2?

   d. If all of (1), (2) and (3) hold, what must be true of the σ_g^2, the n_g, τ^2 and σ^2? What would this imply about the scientific mechanism that causes chlorophyll to be related to phosphorus in lakes?

A scatterplot of log chlorophyll on log phosphorus using unaggregated data is shown for 9 of the 15 lakes in Figure 3 with colors denoting observations from different lakes. These data are a subset of those in Figure 1 and contain less lakes only to maintain clarity by reducing overlap to some extent. Values of fitted regressions (simple linear) to data from individual lakes are presented in Table 2.

3. Comment on whether or not it appears from Figure 3 and Table 2 that the condition you gave in your answer to Question 1 might be met by these data. Do not do any formal test or comparison procedure, just assess the available evidence visually. From an intuitive standpoint, explain why the estimated regression coefficients of Table 1 are so similar for both the original (unaggregated) and the group averaged (aggregated) data.

In Table 1 the coefficient of determination (r^2) is close to 25% greater for the aggregated data than for the original unaggregated data (0.901 versus 0.732). It appears that greater correlations in aggregated data than unaggregated data have been observed in many problems as far back as Karl Pearson in the late 1800s (Knapp 1977). Robinson (1950) outlined the relation between degree of association in aggregated and unaggregated data in the following way. Suppose that $\{(X_{g,i}, Y_{g,i}) : g = 1, \ldots, G; i = 1, \ldots, n_g\}$ are pairs of random variables observed on a total of $N = \sum_{g=1}^{G} n_g$ original sampling units. Let ρ_U denote the ordinary (Pearson product moment) correlation between these N pairs of unaggregated variables, and let ρ_g denote the correlation between pairs of variables within group g. Let $\{\bar{X}_g, \bar{Y}_g) : g = 1, \ldots, G\}$ be group averages of the

Lake	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\sigma}^2$	r^2
45	-1.371	0.914	0.1205	0.222
83	-1.586	1.034	0.3215	0.216
84	0.561	0.641	0.2176	0.088
85	-5.383	2.591	0.3278	0.632
87	-1.988	1.319	0.0874	0.408
88	-3.493	1.652	0.1586	0.242
119	$-.0666$	1.029	0.1070	0.327
131	-0.766	0.878	0.2484	0.025
133	-2.304	1.312	0.1637	0.255
137	5.324	-0.568	0.3932	0.037
139	2.302	0.322	0.1134	0.084
179	1.342	0.513	0.0730	0.128
180	-3.864	1.802	0.4886	0.333
181	-0.590	1.051	0.4088	0.103
185	-0.720	0.951	0.5411	0.234

Table 2: Summary values for simple linear regressions of log chlorophyll on log phosphorus in individual lakes.

$X_{g,i}$ and $Y_{g,i}$ and let ρ_A denote the correlation between these G pairs of averages in the aggregated data. Define the following quantities:

$$\begin{array}{ll} \bar{x} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{g,i} & \bar{x}_g = \frac{1}{n_j}\sum_{i=1}^{n_g} x_{g,i} \\ \bar{y} = \frac{1}{N}\sum_{g=1}^{G}\sum_{i=1}^{n_g} y_{g,i} & \bar{y}_g = \frac{1}{n_g}\sum_{i=1}^{n_g} y_{g,i} \\ SSx_U = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{g,i} - \bar{x})^2 & SSx_A = \sum_{g=1}^{G} n_g(\bar{x}_g - \bar{x})^2 \\ SSy_U = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(t_{g,i} - \bar{x})^2 & SSy_A = \sum_{g=1}^{G} n_g(\bar{y}_g - \bar{y})^2 \\ \eta_x^2 = \frac{SSx_A}{SSx_U} & \eta_y^2 = \frac{SSy_A}{SSy_U} \end{array} \tag{4}$$

Robinson's relation between ρ_U, ρ_A and the collection of ρ_g is,

$$\rho_U = \left[\eta_x^2\eta_y^2\right]^{1/2}\rho_A + \left[(1-\eta_x^2)(1-\eta_y^2)\right]^{1/2} H_w, \tag{5}$$

where $H_w = h(\rho_1, \rho_2, \ldots, \rho_G)$ is a function of the within group correlations.

4. Assume that $H_W \geq 0$ in (5) and let $N = \sum_g n_g$. Show that $0 \leq \eta_x^2 \leq 1$ (and hence, so also is η_y^2).
Hint: You may use without derivation that,

$$SSx_A = \sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{g,i}^2 - N\bar{x}^2$$

$$SSx_U = \sum_{g=1}^{G} n_g\bar{x}_g^2 - N\bar{x}^2.$$

5. Now show that $\rho_A \geq \rho_U$ and that as $\eta_x^2 \to 1$ and $\eta_y^2 \to 1$, ρ_U gets closer to ρ_A (don't worry about the rates at which η_x^2 and η_y^2 are going to 1).

We now turn our attention to the second issue of the effect of transformation of Y_i on what we can learn about the relation between chlorophyll (Y_i) and phosphorus (x_i) in lakes.

6. Some scientists might see that model (1) implies that $E[\log(Y_i)] = \beta_0 + \beta_1 \log(x_i)$ and then infer that the expected value of Y_i is $E(Y_i) = \exp[\beta_0 + \beta_1 \log(x_i)]$ and that the response variables Y_i follow a regression with this expectation function and additive errors. Explain why this is flawed reasoning.

For reasons such as those in your answer to Question 6, we might well wish to develop a regression model for chlorophyll on phosphorus rather than transformations of these quantities. Here, we will leave values of phosphorus on a logarithmic scale to reduce the number of factors to be considered in comparing results with transformed and non-transformed values of chlorophyll. One question in developing a model for a regression of chlorophyll on the logarithm of phosphorus is selection of a random model component. Figure 4 contains a Box-Cox plot for the unaggregated data, which shows a clear linear relation between log means and log standard deviations for binned data. The slope of a straight line fit to the points of Figure 4 is 0.812.

7. Based on the Box-Cox plot of Figure 4 what would you suggest might be the relation between expected values and variances we would desire in a random model component for a regression of chlorophyll on log phosphorus.

8. Two potential random model components that some statisticians would suggest for our desired regression are gamma and lognormal. Parameterize a gamma distribution so that $E(Y) = \alpha/\beta$ and $var(Y) = \alpha/\beta^2$ and a lognormal so that $E(Y) = \exp(\mu + 0.5\sigma^2)$ and $var(Y) = \exp(2\mu + \sigma^2)[\exp(\sigma^2) - 1]$. Note that these moments for a lognormal are what result from $\log(Y) \sim N(\mu, \sigma^2)$ with the usual expression for the normal density. So the random component for a gamma model is that Y_i has probability density function with parameters $\alpha_i > 0$ and $\beta_i > 0$,

$$f(y|\alpha_i, \beta_i) = \frac{\beta_i^{\alpha_i}}{\Gamma(\alpha_i)}\, y^{\alpha_i - 1} \exp(-\beta_i y); \quad y > 0$$

and that for a lognormal model is, for $-\infty < \mu_i < \infty$ and $\sigma_i^2 > 0$,

$$f(y|\mu_i, \sigma_i^2) = \frac{1}{(2\pi\sigma_i^2)^{1/2} y} \exp\left[-\frac{1}{2\sigma_i^2}\left\{\log(y) - \mu_i\right\}^2\right].$$

Formulate the remaining portions of regression models with these two random components if the expectation function is $E(Y_i) = \exp(\eta_i) = \exp[\gamma_0 + \gamma_1 \log(x_i)]$.

9. Suppose the two models from Question 8, which differ only in random component, are fit to the available data. Describe a diagnostic procedure (not necessarily a formal test) that could to used to examine whether there is a difference between the models in terms of their ability to describe the distributions of response variables.

As it turns out, models with gamma and lognormal random components give nearly identical results, and an appropriate diagnostic did not indicate one model as more appropriate than the other. For simplicity we will focus on models with gamma random components.

We now briefly return to the question of the effects of data aggregation, but this time with chlorophyll in its original scale rather than log transformed values. Figure 5 shows a scatterplot of chlorophyll versus log phosphorus for the original data, along with a fitted curve from a model with gamma random component and an exponential regression function. A similar scatterplot and fitted exponential expectation function for aggregated (seasonal mean) chlorophyll and log phosphorus is presented in Figure 6.

Although the fitted expectation functions in Figures 5 and 6 both appear visually reasonable, it is not obvious that the same nonlinear model should be appropriate for both unaggregated and aggregated data. That is, if in a model for the data of Figure 5, $E(Y_{g,i}) = \exp(\eta_{g,i})$ there is no immediate reason that in a model for the data of Figure 6 we would expect $E(\bar{Y}_g) = \exp(\bar{\eta}_g)$. In fact, these cannot both be mathematically true. Suppose that $E(Y_{g,i}) = \exp(\eta_{g,i})$. Aggregating random variables over a group gives,

$$E(\bar{Y}_g) = \frac{1}{n_g}\sum_{i=1}^{n_g} E(Y_{g,i}) = \frac{1}{n_g}\sum_{i=1}^{n_g} \exp(\eta_{g,i}). \tag{6}$$

The question then becomes to what degree this can be approximated as,

$$E(\bar{Y}_g) \approx \exp\left(\bar{\eta}_g\right) = \exp\left[\frac{1}{n_g}\sum_{i=1}^{n_g} \eta_{g,i}\right]. \tag{7}$$

Consider the function on the right hand side of (6. First and second order expansions about the point $(\bar{\eta}_g, \ldots, \bar{\eta}_g)^T$ are,

$$\begin{aligned} \frac{1}{n}\left[\sum_{i=1}^{n} \exp(\eta_{g,i})\right] &\approx \exp(\bar{\eta}_g) \\ \frac{1}{n}\left[\sum_{i=1}^{n} \exp(\eta_{g,i})\right] &\approx \exp(\bar{\eta}_g)\left[1 + \frac{\sum_{i=1}^{n} \eta_{g,i}^2}{2n} - \frac{\bar{\eta}_g^2}{2}\right]. \end{aligned} \tag{8}$$

10. Assuming that the expectation function for the unaggregated data of Figure 5 is correctly specified as $E(Y_{g,i}) = \exp(\eta_{g,i})$, the first order approximation in (8) lends some support to the idea that this expectation function might be a reasonable approximation for the aggregated data of Figure 6 as well. What does the second order approximation in (8) indicate about when this approximation will be more versus less accurate?

Suppose we would now like to develop a hierarchical model for sets of individual lakes including, but not limited to, the ones we have been examining here. Despite the exponential fits across a wide range of values for the logarithm of total phosphorus, suppose that we determine simple straight line expectation functions appear reasonable for individual lakes (this may or may not actually be true, but it simplifies this question to suppose it is true). In a hierarchical model to relate chlorophyll $Y_{g,i}$ to phosphorus or log phosphorus $x_{g,i}$ for lakes $g = 1, \ldots, G$ and observations $i = 1, \ldots, n_g$, the data model would then be,

$$Y_{g,i} = \gamma_{g,0} + \gamma_{g,1} x_{g,i} + \sigma_g \epsilon_{g,i}, \tag{9}$$

where the $\epsilon_{g,i}$ are independent and identically distributed with normal distributions having expected value 0 and variance 1.

Suppose that our intention is to construct a hierarchical model by assigning random parameter distributions to the $\gamma_{g,0}$, $\gamma_{g,1}$ and σ_g^2 for $g = 1, \ldots, G$ and then to further place prior distributions on any parameters that may be involved in these random parameter distributions. We will use $p(x)$ as generic notation to represent the distribution (as a density or mass function) of X, and so forth. Distributions assigned in our model will be

$$\begin{aligned}
\gamma_{g,0} &\sim \text{ iid } p(\gamma_{g,0}|\mu_0, \tau_0^2) \\
\gamma_{g,1} &\sim \text{ iid } p(\gamma_{g,1}|\mu_l \tau_1^2) \\
\sigma_g^2 &\sim \text{ iid } p(\sigma_g^2|\alpha, \beta) \\
\mu_0|\tau^0 &\sim p(\mu_0|\tau_0^2) \\
\mu_1|\tau^1 &\sim p(\mu_1|\tau_1^2) \\
\tau_0^2 &\sim p(\tau_0^2) \\
\tau_1^2 &\sim p(\tau_1^2) \\
\alpha &\sim p(\alpha) \\
\beta &\sim p(\beta)
\end{aligned}$$

The specific forms of these distributions is not important for this question.

11. Using the notation just presented and its extension to joint distributions (e.g., $p(\boldsymbol{\gamma}_0|\mu_0, \tau_0^2)$ is the joint of $\gamma_{g,0}$; $g = 1, \ldots, G$) give the simplest forms possible for the full conditional posterior distributions of μ_0 and σ_g^2. That is, give, for $g = 1, \ldots, G$,

$$\begin{aligned}
p(\mu_0|\cdot) &\propto \\
p(\sigma_g^2|\cdot) &\propto
\end{aligned}$$

12. We are interested in making inference about the chlorophyll-phosphorus regression for lakes in the same region as, but that are not included in our set of 15 sampled lakes. In particular, we would like to make inference about the slope of that unobserved regression, $\gamma_{g^*,1}$ for some lake g^* not included in our set of observed lakes. Suppose we would like to produce a 90% prediction interval for $\gamma_{g^*,1}$. Briefly outline a procedure to compute this quantity.

Figures

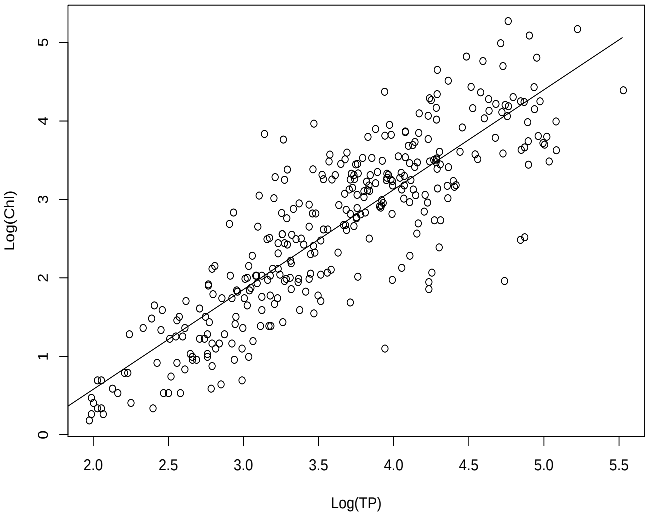


Figure 1: Scatterplot of log chlorophyll on log phosphorus with fitted expectation function from a straight line model.

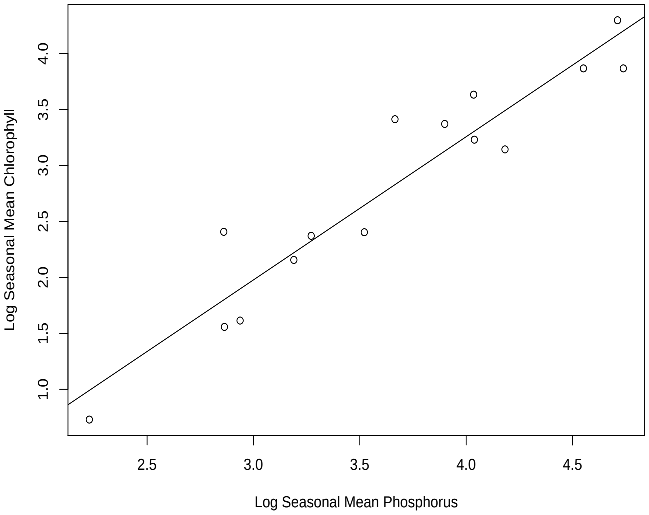


Figure 2: Scatterplot of seasonal mean log chlorophyll on seasonal mean log phosphorus with fitted expectation function from a straight line model.

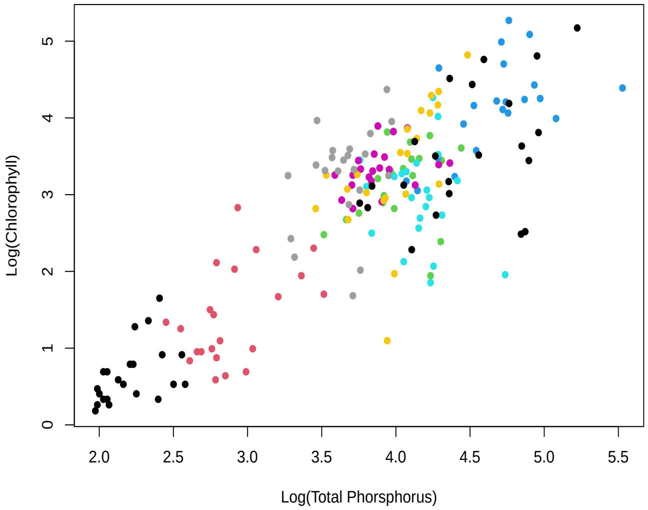


Figure 3: Partial scatterplot of log chlorophyll on log phosphorus for original observations with colors denoting individual lakes.

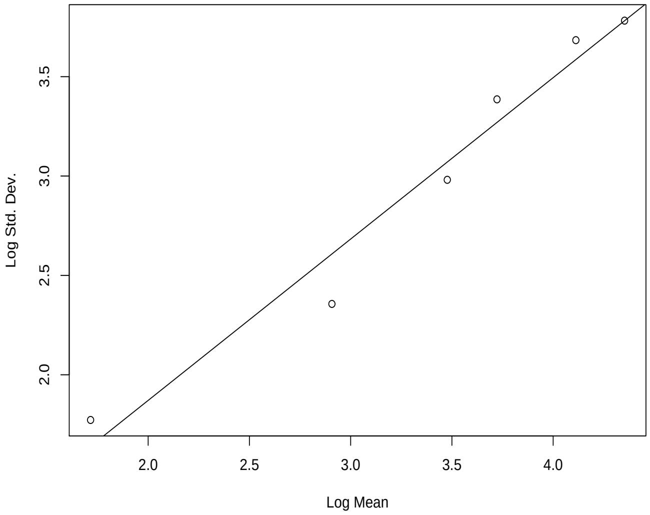


Figure 4: Box-Cox plot for data with chlorophyll in the original scale.

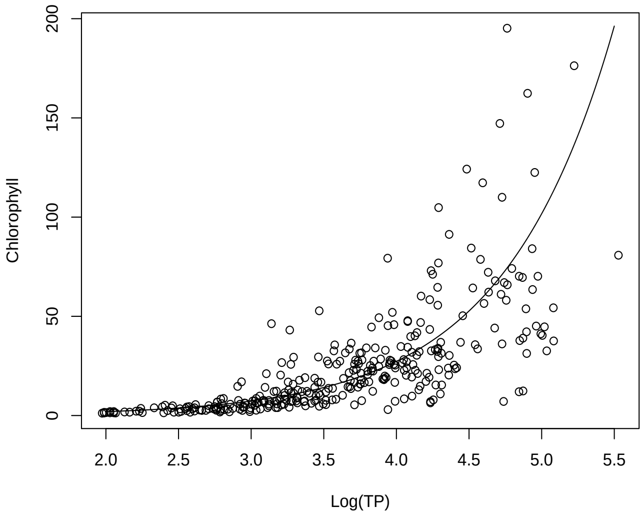


Figure 5: Scatterplot of unaggregated data chlorophyll against log total phosphorus with fitted exponential model.

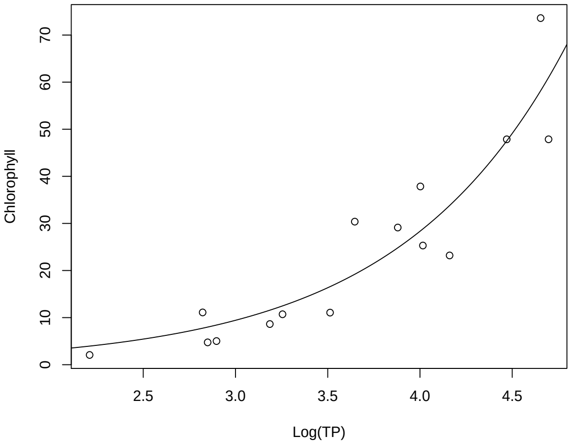


Figure 6: Scatterplot of aggregated data chlorophyll against log total phosphorus with fitted exponential model.

References

Knapp, T.R. (1977), The unit-of-analysis problem in applications of simple correlation analysis to educational research. *Journal of the American Statistical Association* **2**:171-186.

Walker, H.M. (1928). A note on the correlation of averages. *Journal of Educational Psychology* **19**: 636-642.

1. In model (3), if $\beta_{g,0} = \beta_0$ and $\beta_{g,1} = \beta_1$, then the regression coefficients of all three models will be the same and, in particular, $\beta_0 = \alpha_0$ and $\beta_1 = \alpha_1$ in (1) and (2). That is, if (3) is actually,

$$\log(Y_{g,i}) = \beta_0 + \beta_1 \log(x_{g,i}) + \sigma_g \epsilon_{g,i},$$

then

$$\frac{1}{n_g}\sum_{i=1}^{n_g} \log(Y_{g,i}) = \beta_0 + \beta_1 \frac{1}{n_g}\sum_{i=1}^{n_g} \log(x_{g,i}) + \sigma_g \frac{1}{n_g}\sum_{i=1}^{n_g} \epsilon_{g,i}$$

which is the same as,

$$W_g = \beta_0 + \beta_1 z_g + \sigma_g \epsilon_g,$$

where $\epsilon_g = (1/n_g)\sum_i \epsilon_{g,i}$. If, in addition, $\sigma_g = \tau$ for $g = 1, \ldots, G$, then we arrive at exactly model (2) with $\alpha_0 = \beta_0$ and $\alpha_1 = \beta_1$.

2. a. If (3) and (2) both hold, then

$$\frac{\sigma_g^2}{n_g} = \tau^2 \Rightarrow \sigma_g^2 = \tau^2 n_g \text{ for } g = 1, \ldots, G.$$

b. If (3) and (1) both hold, then

$$\frac{\sigma_g^2}{n_g} = \frac{\sigma^2}{n_g} \Rightarrow \sigma_g^2 = \sigma^2 \text{ for } g = 1, \ldots, G.$$

c. If (2) and (1) hold, then

$$\tau^2 = \frac{\sigma^2}{n_g} \Rightarrow n_g = n \text{ for } g = 1, \ldots, G.$$

d. If all of (1), (2) and (3) hold, then

$$\frac{\sigma_g^2}{n_g} = \tau^2 = \frac{\sigma}{n_g} \Rightarrow \tau^2 = \sigma^2 = \sigma_g^2 \text{ and } n_g = n \text{ for } g = 1, \ldots, G.$$

What this would imply about the scientific mechanism underlying the relation between chlorophyll and phosphorus in lakes is that it is manifested in the same way in each situation (lake) in which it occurs. In addition, the observation process is taken to be identical in each situation as well, with observations coming from distributions with possibly unequal means, but the same variance. This implies that variability is caused entirely by the observation or measurement process which is taken to be the same for each lake.

3. No, the condition of Question 1 does not appear to hold for these data. Individual regressions for the different lakes give quite dissimilar estimated coefficients. The reason the regressions of Table 1 are so similar is due to the limited ranges of covariate values that occur within individual lakes. Figure 3 indicates that taking averages of both log chlorophyll and log phosphorus for values within lakes simply “concentrates” the values within small regions of the overall ranges of values, thus leaving the overall pattern similar to that seen in the data of Figure 2.

4. SSx_A will be greater than SSx_U if,

$$\sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{g,i}^2 - N\bar{x}^2 \geq \sum_{g=1}^{n_g} n_g\bar{x}_g^2 - N\bar{x}^2$$
$$\Rightarrow \sum_{g=1}^{G}\sum_{i=1}^{n_g} x_{g,i}^2 \geq \sum_{g=1}^{n_g} n_g\bar{x}_g^2.$$

For each $g = 1, \ldots, G$ we have that

$$\sum_{i=1}^{n_g} x_{g,i}^2 \geq n_g\bar{x}_g^2$$

because

$$\sum_{i=1}^{n_g}(x_{g,i} - \bar{x}_g)^2 = \sum_{i=1}^{n_g} x_{g,i}^2 - n_g\bar{x}_g^2.$$

Then it follows that $SSx_A \geq SSx_U$.

5. Based on Robinson's relation between ρ_U and ρ_A,

$$\rho_U = \left[\eta_x^2\eta_y^2\right]^{1/2}\rho_A + \left[(1-\eta_x^2)(1-\eta_y^2)\right]^{1/2} H_w$$

If $H_w \geq 0$ and with $\eta_x^2 \leq 1$, $\eta_y^2 \leq 1$,

$$\rho_U \leq \left[\eta_x^2\eta_y^2\right]^{1/2}\rho_A. \tag{1}$$

Then, since $\eta_x^2 \leq 1$ and $\eta_y^2 \leq 1$, the result follows that $\rho_U \leq \rho_A$. Directly from (1), as $\eta^2 \to 1$ and $\eta_y^2 \to 1$, ρ_U gets closer to ρ_A.

6. There are two difficulties with the assumption that a simple linear regression for $\log(Y_i)$ implies that $E(Y_i)$ can be found by exponentiating the regression function, both of which have to do with the transformation of a distribution for $\log(Y_i)$ into a distribution for Y_i. First, $\log(\cdot)$ is a concave function so that Jensen's Inequality implies that $E(Y_i) \geq E[\log(Y_i)]$. In fact, if Y_i has a lognormal distribution with parameters μ and σ^2, then $E(Y) = \exp[\mu+\sigma^2/2]$. Secondly, a lognormal distribution is not a location-scale family so that if $\log(Y_i)$ follows an additive error model, then Y_i will not also follow an additive error model.

7. The Box-Cox plot indicates that variances are proportion to expected values raised to a power of twice the Box-Cox slope. That is, if $\mu_i = E(Y_i)$,

$$\text{var}(Y_i) \propto \mu_i^1.6$$

which is between μ_i and μ_i^2. A reasonable place to start is with random components that take $\text{var}(Y_i) \propto \mu_i^2$.

8. To finish formulation of regression models with the two random components given, we need to specify systematic model components to relate $E(Y_i)$ to the covariates x_i, and determine a parameter that will remain constant across observations. Given the indication that we would desire variances to be proportional to squared expected values, the gamma model should be formulated as,

$$\eta_i = \gamma_0 + \gamma_1 x_i$$
$$E(Y_i) = \frac{\alpha}{\beta_i} = \exp(\eta_i)$$
$$\text{var}(Y_i) = \frac{1}{\alpha} E^2(Y_i) \propto E^2(Y_i).$$

For the lognormal model,

$$\begin{aligned}\eta_i &= \gamma_0 + \gamma_1 x_i \\ E(Y_i) &= \exp\left[\mu_i + \frac{1}{2}\sigma^2\right] = \exp(\eta_i) \\ \text{var}(Y_i) &= \exp[2\mu_i + \sigma^2][\exp(\sigma^2) - 1] \propto E^2(Y_i).\end{aligned}$$

9. A diagnostic designed to examine the overall ability of a model to describe the distributions of response variables is based on generalized residuals. For either model, let ψ denote the set of parameters to be estimated and $f_i(y|\hat{\psi})$ the model density using the estimated parameter value, that is, the fitted model. For observed responses $y_1, \ldots, y_n$, generalized residuals are defined as,

$$u_i = \int_{-\infty}^{y_i} f_i(t|\hat{\psi})\, dt$$

If the model leading to $f_i(y|\hat{\psi})$ is a reasonable description of the distribution of response variables, then the generalized residuals computed for that model should resemble a random sample from a uniform distribution on the unit interval. A diagnostic plot comparing the empirical distribution function of the u_i to a uniform $(0, 1)$ distribution function could be used to assess the efficacy of the models. Alternatively, a more formal test for uniform distribution could be conducted with the generalized residuals.

10. While the first-order expansion in (9) provides some justification for the approximation, the second-order expression in (9) indicates that the first order approximation becomes worse as variability in the $\eta_{g,i}$ that compose $\bar{\eta}_g$ increases.
Although not needed as part of the answer, note that since $\eta_{g,i} = \gamma_0 + \gamma_1 x_{g,i}$ an increase in variability (across i) of the $\eta_{g,i}$ implies an increase in the variability of the $x_{g,i}$, which is also related to the range of covariate values in individual lakes. If lakes tend to have small ranges of covariate values (as is the case here), then the approximation would be expected to be reasonably good (as is the case here). If, however, individual lakes had large ranges of covariate values such as would occur if each lake spanned the entire range of values in the entire data set, then we would expect the approximation to be poor.

11. The full conditionals requested are,

$$p(\mu_0|\cdot) \propto \pi(\mu_0|\tau_0^2)\, g(\boldsymbol{\gamma}_0|\mu_0, \tau_0^2)$$
$$p(\sigma_g^2|\cdot) \propto \pi(\sigma_g^2)\, p(\boldsymbol{y}_g|\gamma_{g,0}, \gamma_{g,1}, \sigma_g^2)$$

12. In this setting we would make inferece based on the posterior predictive distribution of the