Two studies were performed to compare the relative effectiveness of two intervention programs designed to reduce drug abuse and prevent behavior associated with high risk of HIV infection. The first study was a small pilot study, and the second study was a much larger study. Participants in both studies all had a history of drug abuse, but they were not using drugs at the beginning of the study in which they participated. The two intervention programs, called Program A and Program B, were compared to assess their relative effectiveness for preventing relapse, i.e., preventing return to drug use. Participants in Program A were required to meet weekly for one year. During that time the participants received instruction to help them recognize "high-risk" situations that may be triggers to relapse (return to drug abuse), and they were taught skills to enable them to cope with such situations without using drugs. Participants in Program B received the same instruction and training as participants in Program A, but they were also required to live together in a large house that was supervised by highly trained counselors (a highly structured communal living arrangement). After completing participation in the assigned treatment program, each participant was monitored for two years to determine if the participant experienced relapse. Part of the focus of these studies is to determine if the highly structured communal living arrangement imposed by Program B has additional benefit for preventing relapse.

Part I

There were 20 subjects available for the pilot study and 10 of those subjects were randomly selected to participate in Program B and the other 10 subjects were assigned to Program A. During the two years of follow-up, 4 of the 10 subjects who completed Program A experienced relapse and none of the 10 subjects who completed Program B experienced relapse.

1. Using a randomization test, what would you conclude about relapse rates from for the two treatment programs?

Part II

Later, a much larger study was done with 400 subjects who were former drug abusers. In this study, 300 of those 400 subjects were randomly selected for enrollment in Program A and the other 100 subjects were assigned to Program B. The results are shown in the following table.

Two-Year Outcome

	Relapsed	Remained Drug Free	Total
Program A	228	72	300
Program B	69	31	100

The "Remained Drug Free" column in the table indicates the number of subjects who remained drug free for two years after completing the assigned treatment program. The "Relapsed" column in the table indicates the number of subjects who experienced at least one relapse within two years after completing the assigned treatment program.

2. Construct and interpret an approximate 95% confidence interval for the difference between the drug free outcome probabilities for the two treatment programs.

3. Recall that one objective is to determine if the highly structured communal living arrangement imposed by Program B has additional benefit for preventing relapse relative to program A. Perform an appropriate test of hypotheses to address this issue. Clearly state the null and alternative hypotheses, give a formula and value for your test statistic, and state your conclusion.

4. An odds ratio can be used to quantify the relative effectiveness of Program B in preventing relapse compared to Program A. The odds of a successful outcome (staying drug free for two years) for Program B is defined as

$$\text{odds of successful outcome for Program B} = \frac{p_B}{1-p_B} \quad ,$$

where p_B denotes the probability of a successful outcome for Program B. Similarly, the odds of a successful outcome for Program A is

$$\text{odds of successful outcome for Program A} = \frac{p_A}{1-p_A}.$$

Then the odds ratio corresponding to the odds of success for Program B divided by the odds of success for Program A is

$$\theta = \frac{\dfrac{p_B}{1-p_B}}{\dfrac{p_A}{1-p_A}}.$$

A point estimate of this odds ratio, $\hat{\theta}$, is obtained by replacing the population proportions, p_A and p_B, with the sample proportions, $\hat{p}_A$ and $\hat{p}_B$, computed from the study data. Construct and interpret an approximate 95% confidence interval for θ from the data given in the table. Provide justification for your formula.

5. Explain how you would assess the actual coverage probability for the method you used to construct the confidence interval for the odds ratio in problem 4 above.

6. The description of the experiment does not indicate how the 400 subjects were recruited for this study. Suppose the 400 subjects were recruited by taking the first 400 men between the ages of 18 and 25 who were arrested for drug possession in Chicago during the recruitment phase of the study. Would this recruitment procedure have any effect on the type of inferences that could be made? Explain.

Part III

Someone suggested that it would have been better to randomly assign the same number of subjects to each treatment program. In part II, 300 subjects were randomly assigned to treatment Program A and 100 were randomly assigned to treatment Program B because it cost three times as much to process a subject through Program B than to process a subject through Program A. Consequently, the researchers spent the same amount of money on each program. For the same total cost the researchers could have processed 150 subjects through Program A and 150 subjects through Program B.

7. With respect to comparing the success rates (drug free rates) for the two programs, explain how you would assess the advantages and disadvantages of randomly assigning 150 subjects to each program relative to randomly assigning 300 subjects to Program A and 100 subjects to program B.

Part IV

Additional information was collected on each participant in the study described in part II. The complete information on each subject consists of the following variables:

$$X_1 = \begin{cases} 0 & \text{if the subject was assigned to Program A} \\ 1 & \text{if the subject was assigned to Program B} \end{cases}$$

$$X_2 = \text{ age at enrollment in a treatment program } - 18 \text{ years}$$

$$X_3 = \text{ subject's race } = \begin{cases} 0 & \text{white} \\ 1 & \text{non-white} \end{cases}$$

$$Y = \begin{cases} 0 & \text{if the subjected relapsed within two years} \\ 1 & \text{if the subject remained drug free for two years} \end{cases}$$

The researchers considered a model for which the conditional distribution of Y was assumed to be binomial with success probability, π, defined by

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_3$$

For this model, "success" is remaining drug free for two years after completing the assigned treatment. R output from fitting this model is shown on page 5 along with a few lines of R code used to generate the output. More questions are presented on page 6.

```
>  # Use the glm function to fit a logistic regression model
>     impact.1 <- glm(Y ~ X1+X2+X3+X1X3,  family=binomial,  data=impact)
>     summary(impact.1)

Deviance Residuals:
     Min          1Q      Median          3Q         Max
-1.0726     -0.7938     -0.6454     -0.1179      1.9323



Coefficients:
                  Estimate  Std. Error   z value     Pr(>|z|)
(Intercept)   -1.77741     0.32881     -5.406     6.46e-08 ***
X1               0.64373     0.29765      2.163     0.03057 *
X2               0.02247     0.01890      1.189     0.23441
X3               0.91912     0.28564      3.218     0.00129 **
X1X3          -1.68658     0.74118     -2.276     0.02287 *
---
Signif. codes:  0    ‘***’  0.001     ‘**’  0.01     ‘*’  0.05     ‘.’  0.1     ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

       Null deviance: 449.87  on 399  degrees of freedom
Residual deviance: 436.56  on 395  degrees of freedom
                      AIC: 446.56



>  # Print the estimated covariance matrix
>     vcov(impact.1)


                   (Intercept)              X1              X2              X3           X1X3
(Intercept)   0.1081175   -0.0386107   -0.0052756   -0.0320362    0.0487058
X1              -0.0386107    0.0885976    0.0005699    0.0303920   -0.0896881
X2              -0.0052756    0.0005699    0.0003572    0.0001248   -0.0012533
X3              -0.0320362    0.0303920    0.0001248    0.0815916   -0.0819860
X1X3           0.0487058   -0.0896881   -0.0012533   -0.0819860    0.5493540
```

8. In the context of this study of the relative effectiveness of treatment programs A and B for helping former drug abusers to stay drug free, interpret the model coefficients β_0, β_1, β_2, β_3, and β_4.

9. Explain how information in the R output displayed on page 5 can be used to test the null hypothesis $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against the alternative that at least one of those parameters is not zero. Obtain a value of your test statistic, if you can, and explain how you would use that value to determine if the null hypothesis should be rejected.

10. Maximum likelihood estimation was used to obtain the values of the parameter estimates shown in the R output displayed on page 5. Explain how the covariance matrix of the parameter estimates, displayed near the bottom of page 5, was produced.

11. Find the maximum likelihood estimate of the probability that a 43-year-old, non-white subject assigned to Program B remains drug free for two years.

12. Using the large-sample normal approximation to the distribution of the parameter estimates, provided by the asymptotic theory for maximum likelihood estimates, show how to construct an approximate 95% confidence interval for the probability that a 43-year-old, non-white subject assigned to Program B remains drug free for two years.

Part I:

1. The objective of the study was to determine if the highly structured living arrangement imposed by Program B reduces the probability of relapse rate relative to Program A. Consequently, the null hypothesis that the relapse rates are the same for the two programs should be tested against the one-sided alternative that the relapse rate is lower for Program B. Based on the random assignment of 10 subjects to each program, the p-value is

$$\frac{\binom{10}{4}\binom{10}{0}}{\binom{20}{10}} = 0.0433$$

Because the p-value is smaller than 0.05, the data provide sufficient evidence to reject the null hypothesis of equal relapse rates for the two programs and conclude that the relapse rate is lower for Program B.

(Note that large sample normal approximation to the null distribution of a two-sample z-statistic should not be used because the number of subjects is too small for the large sample normal approximation to provide an accurate probability.)

Part II:

2. The sample sizes are now large enough to accurately use a large sample normal approximation to the distribution of the difference in the sample proportions to construct an approximate 95% confidence interval. The calculations are as follows:

$$(\hat{p}_B - \hat{p}_A) \pm z_{0.975}\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{300} + \frac{\hat{p}_B(1-\hat{p}_B)}{100}}$$

$$\Rightarrow \quad (0.31-0.24) \pm (1.96)\sqrt{\frac{(0.24)(0.76)}{300} + \frac{(0.31)(0.69)}{100}}$$

$$\Rightarrow \quad (-0.033,\ 0.173)$$

We are 95% confident that the true difference in the true drug-free probabilities for Program B versus Program A is between -0.033 and 0.173.

3. Using p_A to denote the population drug-free probability for Program A and p_B to denote the population drug-free probability for Program B, test the null hypothesis $H_0 : p_A = p_B$ against the one-sided alternative $H_A : p_B > p_A$. Because the estimates of the expected numbers of successes and expected number of failures are both larger than 10 for both programs, the large sample normal

approximation to the null distribution of the two-sample z-statistic can be used with reasonable accuracy. First compute

$$\hat{p} = \frac{72+31}{400} = 0.2575 .$$

Then compute the test statistic

$$z = \frac{(\hat{p}_B - \hat{p}_A)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{300} + \frac{\hat{p}(1-\hat{p})}{100}}} = \frac{(0.31-0.24)}{\sqrt{\frac{(0.2525)(0.7475)}{300} + \frac{(0.2525)(0.7475)}{100}}} = 1.3864 .$$

Because the value of the test statistic is less than 1.645, the null hypothesis cannot be rejected at the 0.05 level of significance. These data do not provide sufficient evidence to conclude that Program B provides a lower relapse probability.

4. Because the natural logarithm of $\hat{\theta}$ has a more nearly symmetrical distribution than $\hat{\theta}$, using the large sample normal approximation for the distribution of $\log(\hat{\theta})$ generally provides a confidence interval with coverage probability closer to the desired level. You can use the delta method to obtain an approximate standard error for $\log(\hat{\theta})$. From the Central Limit Theorem

$$\sqrt{n_A}\,(\hat{p}_A - p_A) \xrightarrow{\text{dist}} N(0,\ p_A(1-p_A)) \quad \text{as } n_A \to \infty$$

and

$$\sqrt{n_B}\,(\hat{p}_B - p_B) \xrightarrow{\text{dist}} N(0,\ p_B(1-p_B)) \quad \text{as } n_B \to \infty .$$

Because $\hat{p}_A$ and $\hat{p}_B$ are independent, it follows that

$$\begin{bmatrix} \sqrt{n_A}\,(\hat{p}_A - p_A) \\ \sqrt{n_B}\,(\hat{p}_B - p_B) \end{bmatrix} \xrightarrow{\text{dist}} N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} p_A(1-p_A) & 0 \\ 0 & p_B(1-p_B) \end{bmatrix} \right)$$

Then it follows from the delta method that

$$\log(\hat{\theta}) = \log\left(\sqrt{n_A}\hat{p}_A\right) - \log\left(\sqrt{n_A}(1-\hat{p}_A)\right) - \log\left(\sqrt{n_B}\hat{p}_B\right) + \log\left(\sqrt{n_B}(1-\hat{p}_B)\right)$$

has a large sample normal distribution with mean $\log(\theta)$ and large sample variance

$$\begin{bmatrix} \frac{1}{p_A}+\frac{1}{1-p_A} \\ \frac{1}{p_B}+\frac{1}{1-p_B} \end{bmatrix}' \begin{bmatrix} \frac{p_A(1-p_A)}{n_A} & 0 \\ 0 & \frac{p_B(1-p_B)}{n_B} \end{bmatrix} \begin{bmatrix} \frac{1}{p_A}+\frac{1}{1-p_A} \\ \frac{1}{p_B}+\frac{1}{1-p_B} \end{bmatrix}$$

$$=\frac{1}{n_A p_A(1-p_A)}+\frac{1}{n_B p_B(1-p_B)}$$

$$=\frac{1}{n_A p_A}+\frac{1}{n_A(1-p_A)}+\frac{1}{n_B p_B}+\frac{1}{n_B(1-p_B)}$$

Then an approximate $(1-\alpha)\times 100\%$ confidence interval for $\log(\theta)$ is

$$\log(\hat{\theta}) \pm z_{1-\alpha/2}\sqrt{\frac{1}{n_A\hat{p}_A}+\frac{1}{n_A(1-\hat{p}_A)}+\frac{1}{n_B\hat{p}_B}+\frac{1}{n_B(1-\hat{p}_B)}}$$

Using the data for this study, an approximate 95% confidence interval for $\log(\theta)$ is

$$\log(1.4227) \pm (1.96)\sqrt{\frac{1}{228}+\frac{1}{72}+\frac{1}{69}+\frac{1}{31}} \quad \Rightarrow \quad (-0.147,\ 0.852)$$

Applying the exponential function to both ends of this interval yields an approximate 95% confidence interval for the odds ratio (0.863, 2.345).

We are 95% confident the odds of remaining drug free using Program B are between 86.3 percent and 235.5 percent of the odds of remaining drug free using Program A.

Scoring note: A few students described how to use bootstrap procedures to construct an approximate 95% confidence interval for the odds ratio. Although they could not actually perform the calculations needed to construct a confidence interval during the exam, this approach received good scores if an appropriate description was provided. The description should contain enough detail so that a knowledgeable statistician could implement it, but there is no need to explicitly talk about the use of any software package or computer code. Then the corresponding solution to problem 5 would need to describe how the coverage probability of the bootstrapped confidence interval could be assessed.

5. Because the construction of the confidence interval in part C is based on a large sample distribution, it may not provide a 95% coverage probability. The coverage probability for this method for constructing a confidence interval can be investigated with simulation methods. Simulate a simple random sample of size 300 from a binomial distribution with success probability 0.76. Simulate an independent random sample of size 100 from a binomial distribution with success probability 0.69. Use the data from these two random samples to evaluate the confidence interval, using the method described in part C. Repeat a large number of times, say 100,000 samples, and estimate the true coverage probability by computing the proportion of simulated confidence intervals that contain 1.4227, the "true" value of the odds ratio for the simulation study.

6. If the 400 subjects who participated in this experiment were obtained by using the first 400 young men between the ages of 18 and 25 who were arrested for drug possession in Chicago, this would not necessarily provide a representative sample male drug users between the ages of 18 and 25 in Chicago. Consequently, inference about the relative effectiveness of Programs A and B could not be made for the entire population of male drug users between the ages of 18 and 25 in Chicago. Inferences would be restricted to the 400 men who participated in the randomized experiment.

Part III:

7. These two studies could be compared by comparing the variances of the difference in the estimated probabilities of remaining drug free for Programs B and A. For the study described in part 2 of this problem, the variance is

$$\sigma_2^2 = \frac{p_B(1-p_B)}{100} + \frac{p_A(1-p_A)}{300}$$

For the study described in part 3 of this problem, the variance is

$$\sigma_3^2 = \frac{p_B(1-p_B)}{150} + \frac{p_A(1-p_A)}{150}$$

The difference in the variances is

$$\begin{aligned}\sigma_2^2 - \sigma_3^2 &= \left[\frac{p_B(1-p_B)}{100} + \frac{p_A(1-p_A)}{300}\right] - \left[\frac{p_B(1-p_B)}{150} + \frac{p_A(1-p_A)}{150}\right] \\ &= \left[\frac{p_B(1-p_B)}{100} - \frac{p_B(1-p_B)}{150}\right] - \left[\frac{p_A(1-p_A)}{150} - \frac{p_A(1-p_A)}{300}\right] \\ &= \frac{p_B(1-p_B)}{300} - \frac{p_A(1-p_A)}{300}\end{aligned}$$

Consequently, the experiment that randomly assigns 150 subjects to each program provides a smaller variance for the difference between the estimated relapse probabilities for the two programs when $p_B(1-p_B) > p_A(1-p_A)$. If $p_B = p_A$, then both experiments provide estimates of the difference in the proportions that have the same variance.

Part IV:

8. In the context of this study of the relative effectiveness of treatment programs A and B for helping former drug abusers to stay drug free, the coefficients in the logistic regression model can be interpreted as follows.

β_0 represents the log-odds that an 18 year old, white subject who completes Program A remains drug free for two years.

Conditioning on the age of the subject, β_1 represents the log-odds that a white subject who completes Program B remains drug free for two years minus the log-odds that a white subject who completes Program A remains drug free for two years. This is the logarithm of a conditional odds ratio.

Conditioning on race and the treatment program, β_2 represents the difference in the log-odds of remaining drug free for two years for subjects who are one year apart in age (older versus younger).

Conditioning on the age of the subject, β_3 represents the log-odds that a non-white subject who completes Program A remains drug free for two years minus the log-odds that a white subject who completes Program A remains drug free for two years. This is the logarithm of a conditional odds ratio.

Conditioning on the age of the subject, β_4 represents difference in log-odds ratios of remaining drug free for two years for program B versus Program A for non-white versus white subjects. This is the logarithm of a ratio of conditional odds ratios.

9. A likelihood test of the null hypothesis $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ against the alternative that at least one of those parameters is not zero is obtained by rejecting the null hypothesis if

$$\text{null deviance} - \text{residual deviance} = 449.87\text{-}436.56 = 13.31 > \chi^2_{4,0.95}$$

10. The covariance matrix of the parameter estimates, displayed near the bottom of page 4, was obtained by first evaluating the negative of the4x4 matrix of expectations of second partial derivatives of the log-likelihood function at the values of the mle's for β_0, β_1, β_2, β_3 and β_4, and then inverting that matrix.

11. The maximum likelihood estimate of the probability that a 43 year old, non-white subject assigned to Program B remains drug free for two years is

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2(25) + \hat{\beta}_3(1) + \hat{\beta}_4(1)(1))}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(25) + \hat{\beta}_3(1) + \hat{\beta}_4(1)(1))}$$

$$= \frac{\exp(-1.77741 + 0.64373 + (0.02247)(25) + 0.91912 - 1.68658)}{1 + \exp(-1.77741 + 0.64373 + (0.02247)(25) + 0.91912 - 1.68658)} = 0.20761$$

12. Define $\underset{\sim}{a}' = [\,1 \quad 1 \quad 25 \quad 1 \quad 1\,]$ and let V denote the estimate of the large sample covariance matrix for the estimated coefficients $\underset{\sim}{\hat{\beta}}$. For these data

$$V = \begin{bmatrix} 0.1081175 & -0.0386107 & -0.0052756 & -0.0320362 & 0.0487058 \\ -0.0386107 & 0.0885976 & 0.0005699 & 0.0303920 & -0.0896881 \\ -0.0052756 & 0.0005699 & 0.0003572 & 0.0001248 & -0.0012533 \\ -0.0320362 & 0.0303920 & 0.0001248 & 0.0815916 & -0.0819860 \\ 0.0487058 & -0.0896881 & -0.0012533 & -0.0819860 & 0.5493540 \end{bmatrix}$$

Then the mle for the probability of remaining drug free is $\hat{\pi} = \dfrac{\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}{1 + \exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}$.

The vector of first partial derivatives of $\hat{\pi} = \dfrac{\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}{1 + \exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}$ with respect to the coefficients is $\hat{\pi}(1 - \hat{\pi})\underset{\sim}{a}$. Using the delta method, a formula for an estimate of the standard deviation of the large sample normal distribution of $\hat{\pi} = \dfrac{\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}{1 + \exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}$ is $\hat{\pi}(1 - \hat{\pi})\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}}$. Consequently, an approximate 95% confidence interval is

$$\frac{\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})}{1 + \exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}})} \pm (1.96)\hat{\pi}(1 - \hat{\pi})\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}}$$

(Note: it was not necessary to evaluate this standard error, but the value is 0.1082 and the approximate 95% confidence interval is

$$0.20761 \pm (1.96)(0.1082) \quad \Rightarrow \quad (-.0045,\ 0.4197)).$$

The previous solution relies on the accuracy of the large sample normal approximation to the distribution of the estimate of $\hat{\pi}$. An alternative way to construct a large sample confidence interval relies on the large sample normal approximation to the distribution of $\underset{\sim}{a}'\underset{\sim}{\hat{\beta}}$, which may be more accurate because $\underset{\sim}{a}'\underset{\sim}{\hat{\beta}}$ may have a more nearly symmetric distribution than $\hat{\pi}$ for smaller samples. The large sample estimate of the standard deviation of $\underset{\sim}{a}'\underset{\sim}{\hat{\beta}}$ is $\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}}$. Then, an approximate 95% confidence interval for $\underset{\sim}{a}'\underset{\sim}{\beta}$ is $\underset{\sim}{a}'\underset{\sim}{\hat{\beta}} \pm (1.96)\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}}$ and an approximate 95% confidence interval for $\pi = \dfrac{\exp(\underset{\sim}{a}'\underset{\sim}{\beta})}{1+\exp(\underset{\sim}{a}'\underset{\sim}{\beta})}$ is

$$\left(\frac{\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}} - (1.96)\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}})}{1+\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}} - (1.96)\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}})}, \frac{\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}} + (1.96)\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}})}{1+\exp(\underset{\sim}{a}'\underset{\sim}{\hat{\beta}} + (1.96)\sqrt{\underset{\sim}{a}'V\underset{\sim}{a}})} \right).$$

It was not necessary to evaluate the endpoints of this interval, but the values are (0.067, 0.488). The relative accuracy of these two approached for constructing an approximate confidence interval can be investigated via simulation.

Part I

Consider the general linear model $\boldsymbol{y} = \boldsymbol{X\beta}+\boldsymbol{\epsilon}$, where $\boldsymbol{y}$ is a random vector, $\boldsymbol{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters, and $\boldsymbol{\epsilon}$ is a vector of random errors with $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ and $\text{Var}(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ for some unknown $\sigma^2 > 0$. Suppose $\text{rank}(\boldsymbol{X}) < p$. We know $\boldsymbol{P_X} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'$ is a symmetric matrix that projects orthogonally onto the column space of $\boldsymbol{X}$ so that $\boldsymbol{P_X X} = \boldsymbol{X}$.

1. Prove that $||\boldsymbol{y} - \boldsymbol{P_X y}||^2 \leq ||\boldsymbol{y} - \boldsymbol{Xb}||^2$ for all $\boldsymbol{b} \in \mathbb{R}^p$.

2. Suppose that $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable. Show that $\text{Var}(\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}) = \sigma^2\boldsymbol{c}'(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{c}$.

3. Find a fully simplified expression for $\text{Var}((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y})$.

4. One way to prove $\boldsymbol{P_X X} = \boldsymbol{X}$ is to first prove that if $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are any matrices satisfying $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_1 = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_2$, then $\boldsymbol{X}\boldsymbol{B}_1 = \boldsymbol{X}\boldsymbol{B}_2$. Prove this result without using $\boldsymbol{P_X X} = \boldsymbol{X}$.

Part II

An experiment was conducted to study the effects of drought stress on two soybean varieties labeled V_1 and V_2. Soybean plants were grown on four benches in a greenhouse. Placed on each bench were 18 pots with one soybean plant per pot. Each plant was treated with one of three watering levels labeled W_1, W_2, and W_3. Watering level W_1 caused plants to experience severe drought stress. Watering level W_2 caused moderate drought stress. Watering level W_3 provided plants with adequate water and resulted in no drought stress. The watering level treatments were delivered by an automated overhead watering system that allowed for plants on one third of each greenhouse bench to be exposed to a specified watering level.

Figure 1 on page 2 depicts the layout of the experiment. Each circle represents a pot containing one soybean plant of the variety indicated by its label (V_1 or V_2). Each dashed and rounded rectangle encloses a set of six plants exposed to a particular watering level indicated by its label (W_1, W_2, or W_3). Each solid rectangle represents a greenhouse bench. Within each bench, the three watering levels were randomly assigned to the three groups of six plants in a completely random way so that all 3! possible assignments were equally likely. The watering level assignments for any one bench were made independently of the watering level assignments for any other bench. Within each group of six plants exposed to a single watering level on a single bench, the arrangement of the variety V_1 plants relative to the variety V_2 plants was randomly selected from the two possible arrangements depicted in Figure 1 on page 2. The two arrangements were equally likely and selected independently for each group of six plants.

At the conclusion of the experiment, the total weight (in grams) of the seeds produced by each soybean plant was recorded. Let $y_{ijk\ell}$ be the total seed weight corresponding to variety V_i, watering level W_j, bench k and plant ℓ ($i = 1, 2$, $j = 1, 2, 3$, $k = 1, 2, 3, 4$, and $\ell = 1, 2, 3$). Various total seed weight averages are provided in Table 1 on page 2.

Figure 1. Experimental layout.

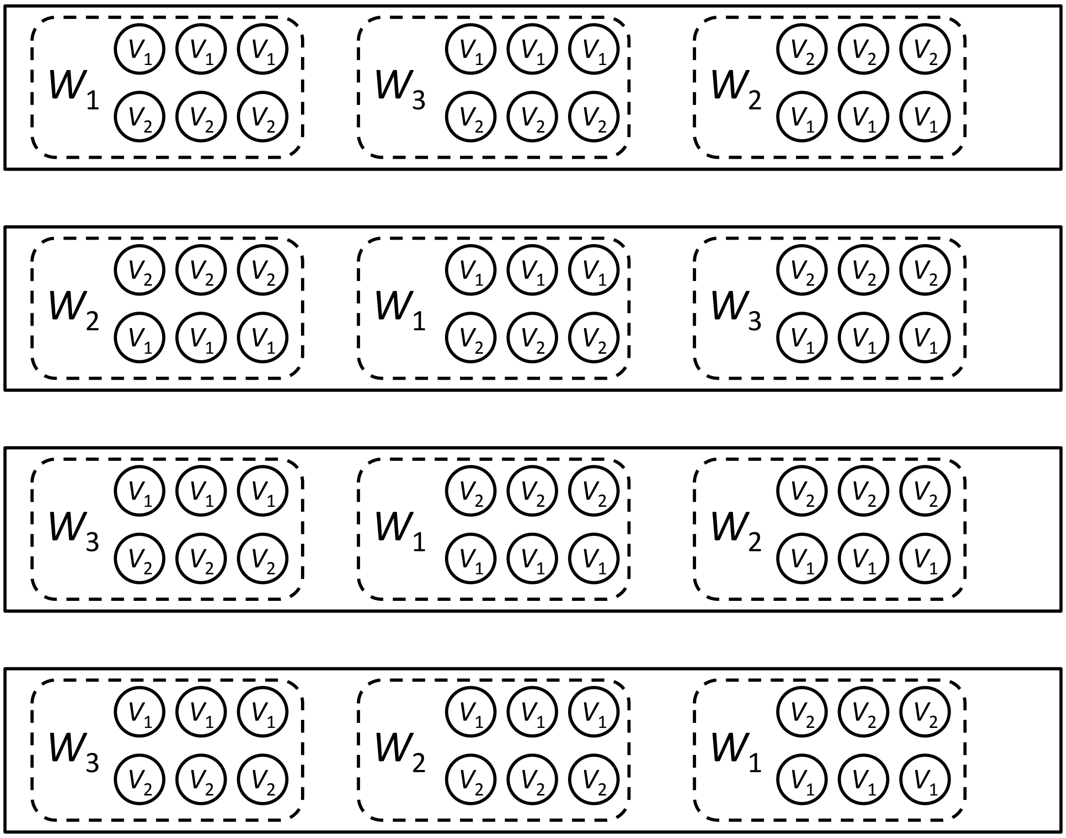

Table 1. Total seed weight averages.

	W_1	W_2	W_3	
V_1	$\bar{y}_{11\cdot\cdot} = 44$	$\bar{y}_{12\cdot\cdot} = 58$	$\bar{y}_{13\cdot\cdot} = 75$	$\bar{y}_{1\cdot\cdot\cdot} = 59$
V_2	$\bar{y}_{21\cdot\cdot} = 32$	$\bar{y}_{22\cdot\cdot} = 48$	$\bar{y}_{23\cdot\cdot} = 85$	$\bar{y}_{2\cdot\cdot\cdot} = 55$
	$\bar{y}_{\cdot 1\cdot\cdot} = 38$	$\bar{y}_{\cdot 2\cdot\cdot} = 53$	$\bar{y}_{\cdot 3\cdot\cdot} = 80$	$\bar{y}_{\cdot\cdot\cdot\cdot} = 57$

Suppose for $i = 1, 2$, $j = 1, 2, 3$, $k = 1, 2, 3, 4$, and $\ell = 1, 2, 3$,

$$y_{ijk\ell} = \mu_{ij} + b_k + g_{jk} + r_{ijk} + e_{ijk\ell}, \tag{1}$$

where μ_{ij} is an unknown parameter and b_k, g_{jk}, r_{ijk}, and $e_{ijk\ell}$ are all random effects. Furthermore, suppose all the random effects are independent and that

$$b_k \sim N(0, \sigma_b^2),\ g_{jk} \sim N(0, \sigma_g^2),\ r_{ijk} \sim N(0, \sigma_r^2), \text{ and } e_{ijk\ell} \sim N(0, \sigma_e^2),$$

where σ_b^2, σ_g^2, σ_r^2, and σ_e^2 are unknown variance components.

5. Find $E(y_{1111})$ in terms of model (1) parameters.

6. Find $\mathrm{Var}(y_{1111})$ in terms of model (1) variance components.

7. Find $\mathrm{Cov}(y_{1111}, y_{1112})$ in terms of model (1) variance components.

8. Using summation notation, provide an expression for the REML estimator of σ_e^2 in terms of the elements of $\{y_{ijk\ell} : i = 1,2,\ j = 1,2,3,\ k = 1,2,3,4,\ \ell = 1,2,3\}$.

9. Show that the REML estimator of σ_e^2 specified in problem **8** is unbiased.

10. Provide an ANOVA table with columns labeled "Source" and "Degrees of Freedom" that indicates how degrees of freedom would be partitioned if model (1) were fit to the data.

Suppose that when model (1) is fit to the data, the REML estimates of the variance components are $\hat{\sigma}_b^2 = 13$, $\hat{\sigma}_g^2 = 10$, $\hat{\sigma}_r^2 = 9$, and $\hat{\sigma}_e^2 = 5$.

11. Compute the value of the F statistic for testing variety main effects.

12. Determine the denominator degrees of freedom for the F statistic in problem **11**.

13. Suppose weather experts predict that plants in the upcoming growing season will experience drought conditions consistent with watering level W_1 with probability 0.1, watering level W_2 with probability 0.3, and watering level W_3 with probability 0.6.

a) For each of varieties V_1 and V_2, estimate the expected total seed weight per plant in the upcoming growing season.

b) Ignoring uncertainty in the predictions of the weather experts, find a 95% confidence interval for the difference between the expected values estimated in problem **13a**.

Researchers collected a simple random sample of approximately one million messenger ribonucleic acid (mRNA) molecules from the root tissue of each soybean plant. They identified the relatively small subset of the mRNA molecules that were produced by *WUE13x*, a gene believed to be involved in water use efficiency. Some of the *WUE13x* mRNA molecules had the correct sequence of nucleic acids while others were missing a key part of the sequence. For variety i, watering level j, bench k, and plant ℓ, let $m_{ijk\ell}$ be the observed number of *WUE13x* mRNA molecules, and let $x_{ijk\ell}$ be the observed number of *WUE13x* mRNA molecules (out of $m_{ijk\ell}$) that were missing the key part of the *WUE13x* sequence. The $m_{ijk\ell}$ and $x_{ijk\ell}$ data were stored in vectors `m` and `x`, respectively, in an `R` workspace, along with vectors `b`, `v`, and `w` that give the levels of bench, variety, and watering level, respectively, for each observation. Use the `R` code and output on pages 5 through 8 to complete the following problems.

14. The researchers want to know if the distribution of the number of mRNA molecules produced by *WUE13x* is the same for all six combinations of soybean variety and watering level considered in their experiment. Conduct one test to answer this question by completing the following parts.

a) Compute the value of a test statistic.

b) State the approximate null distribution of the test statistic (including its degrees of freedom).

c) Provide a p-value.

d) State a conclusion.

15. Compute and interpret an approximate 95% confidence interval that will help the researchers understand how – under severe drought conditions consistent with watering level W_1 – the odds that a *WUE13x* mRNA molecule will be missing a key part of its sequence depend on the soybean variety.

```
> rbind(b,w,v,m,x)
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
b    1    1    1    1    1    1    1    1    1     1     1     1     1
w    1    1    1    1    1    1    2    2    2     2     2     2     3
v    1    1    1    2    2    2    1    1    1     2     2     2     1
m   62   50   61   39   40   35   73   51   67    28    39    31    37
x   21   17   19   30   31   27   32   20   38    24    16    18    15
  [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24]
b     1     1     1     1     1     2     2     2     2     2     2
w     3     3     3     3     3     1     1     1     1     1     1
v     1     1     2     2     2     1     1     1     2     2     2
m    35    34    39    52    33    84    71   115    75    57    53
x    15    15    14    25    18    17    18    26    50    37    39
  [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32] [,33] [,34] [,35]
b     2     2     2     2     2     2     2     2     2     2     2
w     2     2     2     2     2     2     3     3     3     3     3
v     1     1     1     2     2     2     1     1     1     2     2
m   116   144   123    79    81    84    98   146    93   123   123
x    37    42    33    40    32    28    23    46    38    26    33
  [,36] [,37] [,38] [,39] [,40] [,41] [,42] [,43] [,44] [,45] [,46]
b     2     3     3     3     3     3     3     3     3     3     3
w     3     1     1     1     1     1     1     2     2     2     2
v     2     1     1     1     2     2     2     1     1     1     2
m    96    24    22    32    36    25    21    43    33    23    28
x    28    10     8    11    30    15    17    10     5     2    10
  [,47] [,48] [,49] [,50] [,51] [,52] [,53] [,54] [,55] [,56] [,57]
b     3     3     3     3     3     3     3     3     4     4     4
w     2     2     3     3     3     3     3     3     1     1     1
v     2     2     1     1     1     2     2     2     1     1     1
m    31    24    26    30    32    21    23    24    67    35    35
x     4    11     9     6    11     3     8     5     9    13     9
  [,58] [,59] [,60] [,61] [,62] [,63] [,64] [,65] [,66] [,67] [,68]
b     4     4     4     4     4     4     4     4     4     4     4
w     1     1     1     2     2     2     2     2     2     3     3
v     2     2     2     1     1     1     2     2     2     1     1
m    58    52    74    63    46    70    62    53    83    51    69
x    45    40    49     9    10    14    22    21    24     8     8
  [,69] [,70] [,71] [,72]
b     4     4     4     4
w     3     3     3     3
v     1     2     2     2
m    69    69   100    86
x    12     4    12     9
```

```
> b=factor(b)
> w=factor(w)
> v=factor(v)
> o=factor(1:72)
>
> library(lme4)
>
> fullm=glmer(m˜v+w+v:w+(1|b)+(1|b:w)+(1|b:w:v)+(1|o),
+             family=poisson(link = "log"))
> summary(fullm)
Generalized linear mixed model fit by maximum likelihood [’glmerMod’]
 Family: poisson ( log )
Formula: m ˜ v + w + v:w + (1 | b) + (1 | b:w) + (1 | b:w:v) + (1 | o)

      AIC       BIC    logLik  deviance
 596.2365  619.0031 -288.1182  576.2365

Random effects:
 Groups Name        Variance  Std.Dev.
 o      (Intercept) 2.036e-02 1.427e-01
 b:w:v  (Intercept) 2.176e-02 1.475e-01
 b:w    (Intercept) 9.976e-11 9.988e-06
 b      (Intercept) 2.036e-01 4.512e-01
Number of obs: 72, groups: o, 72; b:w:v, 24; b:w, 12; b, 4

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.90002    0.24457  15.946   <2e-16 ***
v2          -0.11977    0.13421  -0.892   0.3722
w2           0.24190    0.13216   1.830   0.0672 .
w3           0.04976    0.13333   0.373   0.7090
v2:w2       -0.17974    0.18864  -0.953   0.3407
v2:w3        0.18855    0.18886   0.998   0.3181
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
      (Intr) v2     w2     w3     v2:w2
v2    -0.272
w2    -0.276  0.502
w3    -0.273  0.498  0.506
v2:w2  0.193 -0.711 -0.701 -0.354
v2:w3  0.193 -0.711 -0.357 -0.706  0.506
```

```
> reducedm=glmer(m~1+(1|b)+(1|b:w)+(1|b:w:v)+(1|o),
+                family=poisson(link = "log"))
> summary(reducedm)
Generalized linear mixed model fit by maximum likelihood [’glmerMod’]
 Family: poisson ( log )
Formula: m ~ 1 + (1 | b) + (1 | b:w) + (1 | b:w:v) + (1 | o)

      AIC       BIC    logLik  deviance
 593.8451  605.2285 -291.9226  583.8451

Random effects:
 Groups Name        Variance  Std.Dev.
 o      (Intercept) 2.042e-02 0.142915
 b:w:v  (Intercept) 3.904e-02 0.197579
 b:w    (Intercept) 2.640e-06 0.001625
 b      (Intercept) 2.009e-01 0.448262
Number of obs: 72, groups: o, 72; b:w:v, 24; b:w, 12; b, 4

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    3.939      0.229    17.2   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
> fullx=glmer(cbind(x,m-x)~v+w+v:w+(1|b)+(1|b:w)+(1|b:w:v)+(1|o),
+             family=binomial(link = "logit"))
> summary(fullx)
Generalized linear mixed model fit by maximum likelihood [’glmerMod’]
 Family: binomial ( logit )
Formula: cbind(x, m - x) ~ v + w + v:w + (1|b) + (1|b:w) + (1|b:w:v) + (1|o)

      AIC       BIC    logLik  deviance
 435.8695  458.6362 -207.9347  415.8695

Random effects:
 Groups Name        Variance  Std.Dev.
 o      (Intercept) 1.994e-02 1.412e-01
 b:w:v  (Intercept) 3.331e-10 1.825e-05
 b:w    (Intercept) 1.010e-01 3.177e-01
 b      (Intercept) 1.180e-01 3.435e-01
Number of obs: 72, groups: o, 72; b:w:v, 24; b:w, 12; b, 4
```

```
Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.94688    0.25486  -3.715 0.000203 ***
v2           2.00094    0.14540  13.761  < 2e-16 ***
w2          -0.02980    0.26354  -0.113 0.909955
w3           0.02955    0.26604   0.111 0.911560
v2:w2       -1.38968    0.19592  -7.093 1.31e-12 ***
v2:w3       -2.24205    0.20040 -11.188  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
      (Intr) v2     w2     w3     v2:w2
v2    -0.267
w2    -0.527  0.258
w3    -0.523  0.256  0.505
v2:w2  0.198 -0.741 -0.352 -0.190
v2:w3  0.194 -0.726 -0.187 -0.360  0.538
>
> reducedx=glmer(cbind(x,m-x)~1+(1|b)+(1|b:w)+(1|b:w:v)+(1|o),
+                family=binomial(link = "logit"))
> summary(reducedx)
Generalized linear mixed model fit by maximum likelihood [’glmerMod’]
 Family: binomial ( logit )
Formula: cbind(x, m - x) ~ 1 + (1 | b) + (1 | b:w) + (1 | b:w:v) + (1 | o)

      AIC       BIC    logLik  deviance
 476.9821  488.3654 -233.4910  466.9821

Random effects:
 Groups Name        Variance Std.Dev.
 o      (Intercept) 0.03796  0.1948
 b:w:v  (Intercept) 0.68992  0.8306
 b:w    (Intercept) 0.07064  0.2658
 b      (Intercept) 0.01040  0.1020
Number of obs: 72, groups: o, 72; b:w:v, 24; b:w, 12; b, 4

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.5644     0.1983  -2.847  0.00442 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Part I

1. For any $\boldsymbol{b} \in \mathbb{R}^P$,

$$\begin{aligned}
||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2 &= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}+\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2 \\
&= (\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}+\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b})'(\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}+\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2+2(\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y})'(\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2+2\{(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{y}\}'(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2+2\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})'\boldsymbol{X}((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2+2\boldsymbol{y}'(\boldsymbol{I}-\boldsymbol{P}_{\boldsymbol{X}})\boldsymbol{X}((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2+2\boldsymbol{y}'(\boldsymbol{X}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{X})((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2+2\boldsymbol{y}'(\boldsymbol{X}-\boldsymbol{X})((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}-\boldsymbol{b}) \\
&= ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2+||\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{b}||^2 \\
&\geq ||\boldsymbol{y}-\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y}||^2.
\end{aligned}$$

2. If $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable, then there exists $\boldsymbol{a}$ such that $\boldsymbol{c}'=\boldsymbol{a}'\boldsymbol{X}$.

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

3. $\mathrm{Var}((\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'(\sigma^2\boldsymbol{I})\boldsymbol{X}\{(\boldsymbol{X}'\boldsymbol{X})^-\}' = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^-\boldsymbol{X}'\boldsymbol{X}\{(\boldsymbol{X}'\boldsymbol{X})^-\}'$.

4.

$$\begin{aligned}
(\boldsymbol{X}\boldsymbol{B}_1-\boldsymbol{X}\boldsymbol{B}_2)'(\boldsymbol{X}\boldsymbol{B}_1-\boldsymbol{X}\boldsymbol{B}_2) &= \boldsymbol{B}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_1+\boldsymbol{B}_2'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_2-\boldsymbol{B}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_2-\boldsymbol{B}_2'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_1 \\
&= \boldsymbol{B}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_2+\boldsymbol{B}_2'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_1-\boldsymbol{B}_1'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_2-\boldsymbol{B}_2'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{B}_1 \\
&= \boldsymbol{0}.
\end{aligned}$$

Because $(\boldsymbol{X}\boldsymbol{B}_1-\boldsymbol{X}\boldsymbol{B}_2)'(\boldsymbol{X}\boldsymbol{B}_1-\boldsymbol{X}\boldsymbol{B}_2)=\boldsymbol{0}$, it follows that $\boldsymbol{X}\boldsymbol{B}_1-\boldsymbol{X}\boldsymbol{B}_2=\boldsymbol{0}$, which implies $\boldsymbol{X}\boldsymbol{B}_1=\boldsymbol{X}\boldsymbol{B}_2$.

Part II

5. $E(y_{1111}) = E(\mu_{11}+b_1+g_{11}+r_{111}+e_{1111}) = \mu_{11}$.

6. $\mathrm{Var}(y_{1111}) = \mathrm{Var}(\mu_{11}+b_1+g_{11}+r_{111}+e_{1111}) = \sigma_b^2+\sigma_g^2+\sigma_r^2+\sigma_e^2$.

7.

$$\begin{aligned}\mathrm{Cov}(y_{1111}, y_{1112}) &= \mathrm{Cov}(\mu_{11} + b_1 + g_{11} + r_{111} + e_{1111}, \mu_{11} + b_1 + g_{11} + r_{111} + e_{1112}) \\ &= \mathrm{Cov}(b_1 + g_{11} + r_{111} + e_{1111}, b_1 + g_{11} + r_{111} + e_{1112}) \\ &= \mathrm{Cov}(b_1, b_1) + \mathrm{Cov}(g_{11}, g_{11}) + \mathrm{Cov}(r_{111}, r_{111}) \\ &= \sigma_b^2 + \sigma_g^2 + \sigma_r^2.\end{aligned}$$

8. $\frac{1}{48}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{4}\sum_{\ell=1}^{3}(y_{ijk\ell} - \bar{y}_{ijk\cdot})^2$

9. For all i, j, k,

$$E\left\{\sum_{\ell=1}^{3}(y_{ijk\ell} - \bar{y}_{ijk\cdot})^2\right\} = E\left\{\sum_{\ell=1}^{3}(e_{ijk\ell} - \bar{e}_{ijk\cdot})^2\right\} = 2\sigma_e^2.$$

Therefore,

$$\begin{aligned}E\left\{\frac{1}{48}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{4}\sum_{\ell=1}^{3}(y_{ijk\ell} - \bar{y}_{ijk\cdot})^2\right\} &= \frac{1}{48}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{4}E\left\{\sum_{\ell=1}^{3}(y_{ijk\ell} - \bar{y}_{ijk\cdot})^2\right\} \\ &= \frac{1}{48}\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{4}2\sigma_e^2 = \frac{1}{48}2\cdot 3\cdot 4\cdot 2\sigma_e^2 = \sigma_e^2.\end{aligned}$$

10. The ANOVA partitioning of the degrees of freedom in this split-plot experiment with multiple observations per split-plot experimental unit is as follows:

Source	Degrees of Freedom
Bench	3
Watering Level	2
Bench × Watering Level	6
Variety	1
Variety × Watering Level	2
Bench × Variety + Bench × Watering Level × Variety	9
Plant(Bench, Watering Level, Variety)	48
C. Total	71

11. $\mathrm{Var}(\bar{y}_{1\cdots} - \bar{y}_{2\cdots}) = \mathrm{Var}(\bar{r}_{1\cdot\cdot} - \bar{r}_{2\cdot\cdot} + \bar{e}_{1\cdots} - \bar{e}_{2\cdots}) = 2\sigma_r^2/12 + 2\sigma_e^2/36 = \sigma_r^2/6 + \sigma_e^2/18$. Thus, $\widehat{\mathrm{Var}}(\bar{y}_{1\cdots} - \bar{y}_{2\cdots}) = \hat{\sigma}_r^2/6 + \hat{\sigma}_e^2/18 = 9/6 + 5/18 = 16/9$, and

$$t = \frac{\bar{y}_{1\cdots} - \bar{y}_{2\cdots}}{\sqrt{\widehat{\mathrm{Var}}(\bar{y}_{1\cdots} - \bar{y}_{2\cdots})}} = \frac{59 - 55}{\sqrt{16/9}} = 3 \Longrightarrow F = 9.$$

12. From the ANOVA table, the term with 9 degrees of freedom that includes Bench × Watering Level × Variety interactions corresponds to the split-plot experimental units because each combination of bench, watering level, and variety is a split-plot experimental unit. Thus, the F-test for split-plot factor main effects has 9 denominator degrees of freedom.

13. **a**) V_1: $0.1 \times 44 + 0.3 \times 58 + 0.6 \times 75 = 66.8$
V_2: $0.1 \times 32 + 0.3 \times 48 + 0.6 \times 85 = 68.6$

b) $\text{Var}(\bar{y}_{1j\cdot\cdot} - \bar{y}_{2j\cdot\cdot}) = \text{Var}(\bar{r}_{1j\cdot} - \bar{r}_{2j\cdot} + \bar{e}_{1j\cdot\cdot} - \bar{e}_{2j\cdot\cdot}) = 2\sigma_r^2/4 + 2\sigma_e^2/12 = \sigma_r^2/2 + \sigma_e^2/6$. Thus, $\widehat{\text{Var}}(\bar{y}_{1j\cdot\cdot} - \bar{y}_{2j\cdot\cdot}) = \hat{\sigma}_r^2/2 + \hat{\sigma}_e^2/6 = 9/2 + 5/6 = 16/3$. It follows that

$$\widehat{\text{Var}}\{0.1(\bar{y}_{11\cdot\cdot} - \bar{y}_{21\cdot\cdot}) + 0.3(\bar{y}_{12\cdot\cdot} - \bar{y}_{22\cdot\cdot}) + 0.6(\bar{y}_{13\cdot\cdot} - \bar{y}_{23\cdot\cdot})\} = (0.1^2 + 0.3^2 + 0.6^2)16/3.$$

A 95% confidence interval for the difference is

$$66.8 - 68.6 \pm 2.26\sqrt{(0.1^2 + 0.3^2 + 0.6^2)16/3} \Longleftrightarrow (-5.34, 1.74).$$

14. **a**) The LRT test statistic is $2(291.9226 - 288.1182) = 7.6088$.

b) χ^2_5

c) `1-pchisq(7.6088,5)` gives a p-value of $0.1791542 \approx 0.18$.

d) The distribution of the number of mRNA molecules produced by *WUE13x* could be the same for all six combinations of variety and watering level considered in this experiment. There is no significant evidence to indicate otherwise.

15. The variety V_2 odds of producing a *WUE13x* mRNA molecule that is missing a key part of its sequence are estimated to be $\exp(2.00094) \approx 7.4$ times the odds for variety V_1. An approximate 95% confidence interval for this multiplicative effect on the odds is $\exp\{2.00094 - 2(0.1454)\} \approx 5.53$ to $\exp\{2.00094 + 2(0.1454)\} \approx 9.89$. The data indicate that – under extreme drought conditions consistent with watering level W_1 – plants of variety V_2 tend to produce malformed *WUE13x* mRNA molecules far more often than plants of variety V_1.

Part I

On 20 April 2010 on the BP-owned Deepwater Horizon oil rig, the largest accidental marine oil spill in the history of the petroleum industry released an estimated 4.9 million barrels of oil into the Gulf of Mexico. Scientists are interested in measuring the effect of the oil spill on vegetation in the Gulf of Mexico. To this end, the scientists obtained vegetation measurements on 4 beaches within the impacted area and 4 beaches outside of the impacted area. Observed differences in vegetation levels on impacted beaches compared to non-impacted beaches may be due to natural differences in the impacted versus non-impacted areas. Thus, from a database constructed **before** the oil spill, the scientists obtained vegetation measurements on 8 other beaches consisting of 4 beaches within the impacted area and 4 beaches outside of the impacted area. Using vegetation measurements from all 16 beaches, the scientific question of interest can be addressed by assessing the interaction between beach type (Non-Impacted vs. Impacted) and period (Before vs. After the oil spill).

For observation i, let Y_i be the vegetation level,

$$I_i = \begin{cases} 1 & \text{if the beach was impacted} \\ 0 & \text{if the beach was not impacted} \end{cases}$$

$$A_i = \begin{cases} 1 & \text{if the measurement was taken after the oil spill} \\ 0 & \text{if the emasurement was taken before the oil spill} \end{cases}$$

Assume

$$y_i \overset{ind}{\sim} N(\mu + \beta_I I_i + \beta_A A_i + \beta_{IA}(I_i \times A_i), \sigma^2) \tag{1}$$

where μ, β_I, β_A, β_{IA}, and σ^2 are unknown model parameters.

1. Provide the cell means in terms of model (1) parameters for each beach-period combination in the format of Table 1.

Table 1: Cell means

	Before	After
Non-impacted		
Impacted		

2. Provide an interpretation for each of μ, β_I, β_A, and β_{IA}. Explain why β_{IA} addresses the scientific question of interest.

3. Model (1) can be written as

$$y = X\beta + \epsilon \tag{2}$$

where $y = (y_1, \ldots, y_{16})^\top$, X is a known matrix of constants, $\beta = (\mu, \beta_I, \beta_A, \beta_{IA})^\top$, and $\epsilon \sim N(0, \sigma^2 \mathrm{I}_{16})$ where I_{16} is an $n \times n$ identity matrix. Give a complete matrix X.

4. Construct a 95% confidence interval for β_{IA} using the numerical facts:

$$X^\top y = (198, 112, 98, 53)^\top,$$

$$y^\top \left(\mathrm{I}_{16} - X[X^\top X]^{-1}X^\top\right)^\top \left(\mathrm{I}_{16} - X[X^\top X]^{-1}X^\top\right) y = 13,$$

$$[X^\top X]^{-1} = \frac{1}{4}\begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 2 & 1 & -2 \\ -1 & 1 & 2 & -2 \\ 1 & -2 & -2 & 4 \end{bmatrix}.$$

Part II

As part of an exploratory data analysis, scientists fit all possible combinations of predictor variables in model (1). The residual sum of squares (RSS) for each model is presented in Table 2.

Table 2: Residual sums of squares for all subset models. 'Yes' indicates that term was included in the model while '–' indicates the term was not included.

Model	I	A	$I \times A$	RSS
1	–	–	–	66
2	–	–	Yes	61
3	–	Yes	–	65
4	–	Yes	Yes	56
5	Yes	–	–	19
6	Yes	–	Yes	15
7	Yes	Yes	–	19
8	Yes	Yes	Yes	13

5. Determine the residual degrees of freedom for each model in Table 2.

6. Perform an F-test to test $H_0 : \beta_{IA} = 0$ versus $H_1 : \beta_{IA} \neq 0$ in the context of the full model. Be sure to include

a) the value of the test statistic,

b) its distribution under the null hypothesis,

c) the p-value for the test, and

d) a scientific conclusion.

Part III

Reviewers of the scientists' manuscript suggested that model uncertainty should be taken into account. One approach involves the use of BIC (Bayesian information criterion) where BIC is defined, up to a constant, as

$$\text{BIC} = -2 \log L(\hat{\theta}) + k \log(n), \tag{3}$$

where $L(\hat{\theta})$ is the likelihood evaluated at the MLE of the model parameter vector θ, k is the dimension of the parameter space, and n is the number of observations.

7. Provide a justification for why BIC can be useful for comparing competing models.

8. In the standard multiple linear regression setting, BIC can be written (aside from a constant) as

$$\text{BIC} = n \log(RSS/n) + k \log(n).$$

Derive this case of BIC.

9. BIC model probability for model j can be calculated using the formula

$$\frac{P(M_j)e^{-\text{BIC}_j}}{\sum_\ell P(M_\ell)e^{-\text{BIC}_\ell}}$$

where $P(M_j)$ is the prior model probability for model j. Assuming a uniform prior over models, calculate BIC and BIC model probabilities for all models in Table 2.

10. To determine the marginal inclusion probability for an interaction term, the BIC model probabilities are summed for those models that include the interaction term. Calculate the marginal inclusion probability for the interaction term.

11. Is the scientific conclusion based on this marginal inclusion probability consistent with the scientific conclusion based on the F-test in problem **6**? Why or why not?

Part IV

BIC model probabilities are approximations to fully Bayesian posterior model probabilities. In order to construct a fully Bayesian posterior model probability, we need 1) priors for model parameters within each model and 2) a prior over models.

For the following assume a joint prior for β and σ^2 in model j given by

$$\pi_j(\beta, \sigma^2) \propto (\sigma^2)^{-p/2} e^{-(\beta-m_j)^\top C_j^{-1}(\beta-m_j)/2\sigma^2} \left(\frac{1}{\sigma^2}\right)^{a+1} (\sigma^2)^{-a-1} e^{-b/\sigma^2}. \tag{4}$$

12. In these regression models, the *prior predictive distribution* under model M_j is

$$p(y|M_j) = \int_{\mathbb{R}^+} \int_{\mathbb{R}^k} f(y|\beta, \sigma^2)\pi_j(\beta, \sigma^2) d\beta d\sigma^2$$

where $f(y|\beta, \sigma^2)$ is the conditional density for the data y given the parameters β and σ^2. Derive the prior predictive distribution for a generic model j.

13. When this prior predictive distribution is evaluated at the observed data y, the resulting value is called the *marginal likelihood*. State the formula for calculating the posterior model probability for model j, i.e. $P(M_j|y)$, using these marginal likelihoods and a prior over models that assigns probability $P(M_j)$ to model j.

14. In an attempt to be uninformative in prior specification, the scientist chooses $m_j = 0$, $C_j = 10^9 \mathrm{I}_k$, $a = 1$, and $b = 1$ for the prior in equation (4). Table 3 provides the logarithm of the marginal likelihoods for each of the models in Table 2 using this prior. Calculate the posterior model probabilities for the models in Table 3 assuming a uniform prior over the models.

Table 3: Log marginal likelihood for all subset models

Model	I	A	$I \times A$	$\log p(y\|M_j)$
1	–	–	–	–48
2	–	–	Yes	–58
3	–	Yes	–	–59
4	–	Yes	Yes	–68
5	Yes	–	–	–48
6	Yes	–	Yes	–57
7	Yes	Yes	–	–59
8	Yes	Yes	Yes	–66

15. Calculate the sum of the posterior model probabilities for models with the interaction.

16. Is the scientific conclusion based on these posterior model probabilities consistent with the conclusions based on the BIC marginal inclusion probabilities in problem **9** and the F-test in problem **6**? Why or why not?

Part I

1. Cell means table:

Table 1: Cell means

	Before	After
Control	μ	$\mu + \beta_A$
Impacted	$\mu + \beta_I$	$\mu + \beta_I + \beta_A + \beta_{IA}$

2.
   - μ is the mean vegetation level in the control beaches before the oil spill
   - β_I is the increase in mean vegetation level before the oil spill of control to impacted beaches
   - β_A is the increase in mean vegetation level of control beaches from before to after the oil spill
   - β_{IA} is the difference from the additive impact of moving from control to impacted and before to after the oil spill

   If $\beta_{IA} = 0$, then the mean vegetation level in impacted beaches after the oil spill can be explained by differences between control and impacted beaches and differences in regional climatic changes before and after the oil spill. Thus a negative β_{IA} indicates a negative impact on mean vegetation levels associated with the oil spill after accounting for differences between control and impacted beaches and other differences before and after the oil spill.

3. If the first 4 observations are non-impacted beaches measured before the oil spill, the next 4 observations are non-impacted beaches measured after the oil spill, the next 4 observations are impacted beaches measured before the oil spill, and the final 4 observations are impacted beaches measured after the oil spill, the the design matrix is given below. Row permutations of this matrix are the other possible answers.

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

4. $\hat{\beta}_{IA}$ is the 4th element of $[X^\top X]^{-1}X^\top y$, so we need

$$\hat{\beta}_{IA} = \frac{1}{4}(1, -2, -2, 4) \times (198, 112, 98, 53)^\top = -2.5$$

The standard error is $\sqrt{\epsilon^\top \epsilon/(n-k)}$ where k is the number of regression coefficients and $\epsilon = \left(I - X^\top[X^\top X]^{-1}X^\top\right) y$, so we have

$$\sqrt{\frac{\epsilon^\top \epsilon}{n-k}} = \sqrt{\frac{13}{16-4}} \approx 1.04.$$

The 95% confidence interval is

$$\hat{\beta}_{IA} \pm t_{12}(.975) \times 1.04 = -2.5 \pm 2.18 \times 1.04 = (-4.77, -0.23)$$

Part II

5. The degrees of freedom is just the number of observations (16) minus the number of predictor variables in the model, i.e. the number of 'Yes's in that row plus 1 (for the intercept).

Model	Degrees of freedom
1	15
2	14
3	14
4	13
5	14
6	13
7	13
8	12

6. We need to compare models 7 and 8.
   - The value of the test statistic is
     $$F = \frac{(19-13)/(13-12)}{13/12} \approx 6.$$
   - The distribution under the null hypothesis is an F distribution with 1 numerator and 12 denominator degrees of freedom.
   - The pvalue is 0.0306218.
   - We reject the null hypothesis of no interaction between before/after and control/impact, i.e. we reject the null hypothesis of no impact of the oil spill after adjusting for differences between control and impacted sites as well as external differences between sites before and after the oil spill.

Part III

7. Model criterion such as BIC balance the fit of the model, measured by $-2\log L(\hat{\theta})$, with additional complexity due to additional parameters in the model, measured by k.

8. Note that the MLE for $\hat{\sigma}^2$ is $\hat{\sigma}^2 = (y - X\hat{\beta})^\top(y - X\hat{\beta})/n = RSS/n$.

$$\begin{aligned} L(\hat{\theta}) &= (2\pi\hat{\sigma}^2)^{-n/2}e^{-(y-X\hat{\beta})^\top(y-X\hat{\beta})/2\hat{\sigma}^2} \\ \log L(\hat{\theta}) &= -n/2\log(2\pi) - n \times \log(\hat{\sigma}^2)/2 - (y - X\hat{\beta})^\top(y - X\hat{\beta})/2\hat{\sigma}^2 \\ &= C - n\log(RSS/n)/2 \\ \text{BIC} &= n\log(RSS/n) + k\log(n) \end{aligned}$$

where $C = -n/2[\log(2\pi) - 1]$ is constant for all models and thus cancels when comparing BIC values.

9. BIC and BIC model probabilities are provided in the table below.

Table 2: BIC and BIC model probabilities (BMP)

Model	BIC	BMP
1	64	0.00
2	60	0.00
3	61	0.00
4	56	0.00
5	42	0.00
6	35	0.01
7	39	0.00
8	30	0.99

10. The marginal inclusion probability is approximately 1.

11. Yes, this is consistent with the results from F-test in that we believe the interaction term is important, i.e. the oil spill resulted in statistically significant change in mean vegetation level.

Part IV

12. The key is to recognize that this prior is a normal-gamma distribution, i.e.

$$\beta|\sigma^2 \sim N(m_j, \sigma^2 C_j) \quad \sigma^2 \sim IG(a, b).$$

Since $y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 \mathrm{I}_n)$, we have

$$y|\sigma^2 \sim N(Xm_j, \sigma^2[XC_jX^\top + I_n])$$

when integrating out β and

$$y \sim t_{2a}(Xm_j, b[XC_jX^\top + I_n]/a)$$

when integrating out σ^2. That is, the prior predictive distribution is a multivariate t distribution with $2a$ degrees of freedom, location vector Xm_j, and scale matrix $b[XC_jX^\top + I_n]/a$.

13. To calculate the posterior probability of a model, use

$$p(M_j|y) = \frac{p(y|M_j)p(M_j)}{\sum_\ell p(y|M_\ell)p(M_\ell)}.$$

14. The posterior model probabilities are provided in the table below.

Table 3: Log marginal likelihood for all subset models

Model	I	A	IA	Log p(y)	Posterior probability
1	–	–	–	-48	0.63
2	–	–	Yes	-58	0.00
3	–	Yes	–	-59	0.00
4	–	Yes	Yes	-68	0.00
5	Yes	–	–	-48	0.37
6	Yes	–	Yes	-57	0.00
7	Yes	Yes	–	-59	0.00
8	Yes	Yes	Yes	-66	0.00

15. The sum of the posterior model probabilities for models with an interaction is approximately 0.

16. No, this result contradicts previous results based on the F-test and BIC marginal inclusion probabilities. This is an example of Lindley's Paradox. The reason is because the extremely vague prior for the model parameters, e.g. $\pi_j(\beta, \sigma^2)$, penalizes the fully Bayesian posterior probabilities via the integration that is required to calculate the prior predictive distribution. Since the models that have more parameters are integrating over a larger dimensional space, the penalty is larger and, in this case, overwhelmes the benefit of including additional parameters.