Corn yields in the US have increased dramatically over the years. For example, the average corn yield increased from 40 bushels / acre in 1950 to 177 bushels / acre in 2021. Plant breeding, i.e., selecting genetic lines with higher yield, is one of many reasons for the yield increase. Others include changes in weather, increased fertilization, better weed control, and growing corn at a higher density (# plants / acre). The data for this problem are based on recent ISU studies designed to estimate how much of the increase in yield can be attributed to plant breeding. This is called genetic gain and is estimated by collecting corn genotypes released in different years, planting them in the same year, at the same site, at the same density, and with the same amounts of fertilization and weed control. Any observed differences among the genotypes can then be attributed to genetic gain. The dataset for this question has 11 genotypes, released in 9 different years, from 1983 to 2017. Two genotypes were released in 1986 and two were released in 2006. The different parts of this question use different subsets of the data.

Show your work for all questions.

Part I. The first set of analyses use data from one site. The experimental design at this site was a randomized complete block design, with 2 blocks. One plot was damaged by hail, so there are data from only 21 plots.

1. Consider model (1)

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

where μ is a common intercept, α_i is the block effect for block i, and τ_j is the genotype effect for genotype j. Here are error sums-of-squares (SSE) and error degrees of freedom (dfE) for model 1 and possible simplifications:

Model	SSE	dfE
E $Y_{ij} = \mu$	2325.9	20
E $Y_{ij} = \mu + \alpha_i$	2134.5	19
E $Y_{ij} = \mu + \tau_j$	1193.2	10
E $Y_{ij} = \mu + \alpha_i + \tau_j$	901.9	9

Compute the type III F-statistic that tests the null hypothesis that $\tau_1 = \tau_i, \forall i = 2, \ldots, 11$.

2. Based on the F-statistic, will the p-value associated with the test in **Question 1**. be > 0.05 or ≤ 0.05? Briefly explain your answer.

Figure 1 is a plot of Y (marginal means for each genotype) against T (their year of release). There are 2 years (1986 and 2006) in which two genotypes were released.

Based on the pattern shown in Figure 1, you consider model (2), which is fit to the data from all 21 plots.

$$\begin{aligned} Y_{ij} &= \mu + \alpha_i + T_j\beta + \varepsilon_{ij} \\ \varepsilon_{ij} &\overset{iid}{\sim} N(0, \sigma_r^2), \end{aligned} \tag{2}$$

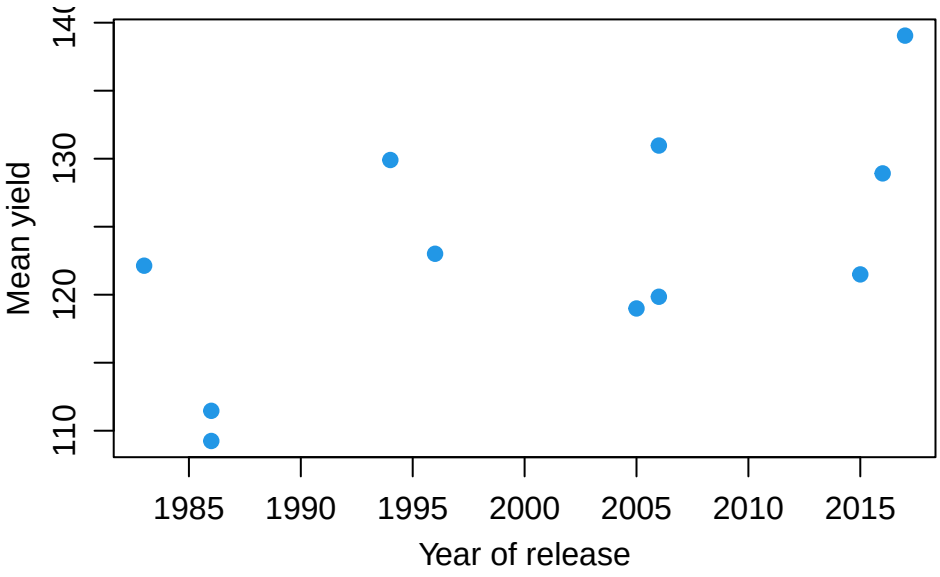

Figure 1: Mean yield for each genotype as a function of year of release.

where T_j is the year of release for genotype j.

3. Give a careful interpretation of β suitable for a plant breeder.

4. Give a careful interpretation of σ_r^2. Your interpretation should include a description of the sources of variability included in σ_r^2.

Both model (1) and model (2) can be written in matrix form, $\boldsymbol{Y} = \boldsymbol{X}_k\boldsymbol{\beta}_k + \boldsymbol{\varepsilon}$, where k specifies model (1) or (2). The first few rows of $\boldsymbol{X}_1$ are:

μ	α_1	α_2	τ_1	τ_2	$\cdots$	τ_{11}
1	1	0	1	0	$\cdots$	0
1	0	1	1	0	$\cdots$	0
1	1	0	0	1	$\cdots$	0

The first few rows of $\boldsymbol{X}_2$ are:

μ	α_1	α_2	T_j
1	1	0	1996
1	0	1	1996
1	1	0	2005

Define $\boldsymbol{P}_k = \boldsymbol{X}_k(\boldsymbol{X}_k^T\boldsymbol{X}_k)^-\boldsymbol{X}_k^T$, where $\boldsymbol{A}^-$ is a generalized inverse of matrix $\boldsymbol{A}$.

5. What name is commonly used for the vector of values computed as $\boldsymbol{P}_k\boldsymbol{Y}$?

6. In terms of quantities used in an ANOVA, what is $\boldsymbol{Y}^T(\boldsymbol{I}-\boldsymbol{P}_1)\boldsymbol{Y}$?

7. In terms of quantities used in an ANOVA, what is $\boldsymbol{Y}^T(\boldsymbol{P}_1-\boldsymbol{P}_2)\boldsymbol{Y}$?

8. If $\boldsymbol{P}_2 \neq \boldsymbol{P}_1$, does a test based on $\boldsymbol{Y}^T(\boldsymbol{P}_1-\boldsymbol{P}_2)\boldsymbol{Y}$ answer a useful question about the appropriateness of model 2? If so, explain what question is being answered.

Part 2. The density (# seeds per acre) at which corn is planted has increased over the years. In the 1980s corn was grown at 20,000 seeds/acre (20K). In the 1990s, that increased to 25,000 seeds per acre (25K), and from 2000 to now, has been 35,000 seeds per acre (35K). The analyses in part I compared genotypes planted at the same density, 35K. The analyses in part 2 consider both genotype and planting density. Each of the two blocks included additional plots for five genotypes at the standard density for when they were released. The investigators only consider the 6 genotypes planted at more than one density. Table 1 gives the number of plots and year of release (YOR) for each combination of genotype and planting density used in this part.

Genotype	Year Of Release	Planting Density 20K	25K	35K
A	1983	2		2
B	1986	2		2
C	1986	2		2
D	1994		2	2
E	1996		2	2
F	1997		2	2

Table 1: Year of release for each genotype and the number of plots planted at each density. Blanks indicate that density was not used for that genotype.

Initially consider the model (3):

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \tau_j + \delta_k + \tau\delta_{jk} + \varepsilon_{ijk}, \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2_{gd}) \end{aligned} \tag{3}$$

where μ, α_i, and τ_j are as defined in model (1), δ_k is the density effect for density k, and $\tau\delta_{jk}$ is the genotype by density interaction effect.

9. Fill in the degrees of freedom (df) in the following ANOVA table. There are a total of 24 observations. Note: G*D is the Genotype*Density interaction

Source	df
Block	
Genotype	
Density	
G*D	
Error	

10. Carefully describe the null hypothesis tested by the F test of the Genotype*Density interaction for these data.

11. The p-value associated with Block is 0.53 so the investigators consider dropping this term from the model. Is it appropriate to drop the block term from the model? Briefly explain why or why not.

The investigators decide to drop the Genotype*Density term from the model because its p-value is 0.74. The model used for the rest of the questions in this part is

$$\begin{aligned} Y_{ijk} &= \mu + \alpha_i + \tau_j + \delta_k + \varepsilon_{ijk} \\ \varepsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma_d^2), \end{aligned} \tag{4}$$

where the terms are the same as in model (3). Table 2 gives means for each observed combination of genotype and density.

Genotype	Planting Density		
	20K	25K	35K
A	123		116
B	125		109
C	126		110
D		132	129
E		139	125
F		129	125

Table 2: Average yield for each observed combination of genotype and density. Blanks indicate that density was not used for that genotype.

Table 3 gives parts of the ANOVA table for model (4).

Source	MS	F
Block	25.11	0.52
genotype	183.5	3.80
density	323.1	6.69
residual	48.3	

Table 3: Sources, mean squares (MS), and F statistics for model (4).

12. Define δ_{20} and δ_{25} as the effects associated with densities 20K and 25K, respectively. You want to estimate $\delta_{20} - \delta_{25}$, the mean change in yield for a genotype if the planting density is increased from 20K to 25K. If this quantity is not estimable, say so. If it is estimable:

a) estimate $\delta_{20} - \delta_{25}$

b) calculate the standard error of that estimate.

13. The genotypes grown at 20K were released earlier than those grown at 25K. Is this a concern with interpreting $\delta_{20} - \delta_{25}$ as the change in yield due to increasing density from 20K to 25K? Briefly explain your answer.

Part 3: We return to considering only one planting density, 35K. The experiment in part 1 was repeated at 4 sites. Two sites were in very productive (fertility = high) soils; two were in less productive (fertility = low) soils. The experimental design at each site is a randomized complete block design with 2 blocks. The same 11 genotypes are used at all sites. Two plots were damaged by hail so there are a total of 86 observations.

14. Here is part of an ANOVA table for an appropriate analysis of these data. Some sources of variability customarily included in an appropriate analysis are missing. Fill in missing sources of variability, the degrees for freedom (df) for all sources, and whether a term should be considered a fixed or a random effect.

Source	df	Fixed or Random
Site(Fertility)		
Genotype*Fertility		
Error		

15. The mean squares (MS) and expected mean squares (EMS) for 3 sources of variability are:

Source	MS	EMS
Site(Fertility)	22.55	$\sigma^2_{error} + 21.333\ \sigma^2_{site}$
Genotype*Fertility	116.89	$\sigma^2_{error} + Q(G * F)$
Error	169.78	σ^2_{error}

Note: $Q(G * F)$ is a quadratic form involving the genotype*fertility interaction effects.

Calculate ANOVA based estimates of σ^2_{error} and σ^2_{site}.

Part I: Genotypes and year of release

1. This is a model comparison between E $Y_{ij} = \mu + \beta_i + \tau_j$ and E $Y_{ij} = \mu | \beta_i$. MS$_{genotypes} = (2134.5 - 901.9)/(19 - 9) = 1232.6/10 = 123.3$. MS$_{error} = 901.9/9 = 100.2$, so $F = 123.2/100.2 = 1.23$

2. > 0.05. 0.95 quantiles for F distributions depend on the numerator and denominator df, but they are commonly > 2. When one or both df are small, the 0.95 quantiles are larger than 2 and often very much so. 1.23 is smaller than 2.

3. β is the average difference in yield between a genotype released in year Y and one released in year $Y + 1$ when compared in the same block.

4. σ_r^2 is the variance of observations (genotype in a block) around the regression line. This includes:
variability between genotype-specific observations not accounted for by block effects,
variability of genotype means around the block-specific regression line.

5. $\boldsymbol{P}_1\boldsymbol{Y}$ is the vector of predicted values using model 1.

6. $\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P}_1)Y$ is the error sums-of-squares associated with fitting model 1

7. $\boldsymbol{Y}^T(\boldsymbol{P}_1 - \boldsymbol{P}_2)Y$ is the change in error sums-of-squares associated with fitting model 1 after fitting model 2.
Or: the model sums-of-squares for model 1 after fitting model 2.

8. Yes, this assesses the lack of fit of the regression. It is the numerator sum-of-squares in an ANOVA lack-of-fit test.

Part 2: Genotypes and planting density

9.

Source	df
Block	1
Genotype	5
Density	2
G*D	4
Error	11

Note: I am especially looking for the G*D interaction df. It is not 2×5 = 10 because of the missing combinations of genotype and density.

10. The null hypothesis has two parts: The difference between 20K and 35K is the same for genotypes A, B, and C AND The difference between 25K and 35K is the same for genotypes D, E, and F.
Note: Writing the null hypothesis requires some care because of the missing cells.

11. No. Blocks are part of the randomization scheme and can not be dropped from the model.

12. Yes, it is estimable.

a)

$$\begin{aligned}\hat{\delta}_{20} - \hat{\delta}_{25} &= \left(\overline{Y}_{A,B,C,\,20} - \overline{Y}_{A,B,C,\,35}\right) - \left(\overline{Y}_{D,E,F,\,25} - \overline{Y}_{D,E,F,\,35}\right) \\ &= (124.7 - 111.7) - (133.3 - 126.3) \\ &= 13.0 - 7.0 = 6.0\end{aligned}$$

b) Each sample mean in Table 2 has a standard error of $\sqrt{\hat{\sigma}_d^2/2} = \sqrt{48.3/2} = 4.91$. All the estimated sample means are independent, so the variance of the desired contrast is $4.91 * \sqrt{4/3} = 5.67$

13. No, because the appropriate estimate is a comparison among densities within each genotype. OR: No, but the comparison between densities in model (**??**) is adjusted for differences between genotypes.

Part III: Repeated at multiple sites

14.

Source	df	Fixed or Random
Fertility	1	Fixed
Site(Fertility)	2	Random
block(Site, fertility)	4	Fixed or Random
Genotype	10	Fixed
Genotype*Fertility	10	Fixed
Residual	58	Random

15.

$$\begin{aligned}\hat{\sigma}^2_{error} &= MS_{Residual} \\ \hat{\sigma}^2_{site} &= (MS_{site(fertility)} - MS_{error})/21.33 \\ &= (22.55 - 169.8)/21.33 \\ &= -6.90\end{aligned}$$

Part I The "Children Ever Born" (CEB) dataset consists of grouped data on the number of births of Fijian women. The women are described according to their marriage duration in years in ordinal levels: (0-4, 5-9, 10-14, 15-19, 20-24, 25-29); their place of residence (Suva, Urban, or Rural); and, their level of education (none, lower primary, upper primary, secondary or greater). The count, mean, and variance of the number of children ever born, and the group size, is given for each group of women by cross-classified factorial level. These summaries are sufficient to model counts of children ever born by a Poisson distribution (each individual woman's count is not needed).

The following `R` code produces the snapshot of CEB data displayed in Table 1.

```
ceb <- read.table('ceb.dat')
head(ceb)
```

##		dur	res	educ	mean	var	n	Y
##	1	0-4	Suva	none	0.5	1.14	8	4
##	2	0-4	Suva	lower	1.14	0.73	21	24
##	3	0-4	Suva	upper	0.9	0.67	42	39
##	4	0-4	Suva	sec+	0.73	0.48	51	37
##	5	0-4	urban	none	1.17	1.06	12	14
##	6	0-4	urban	lower	0.85	1.59	27	23

Table 1: First five rows of CEB data.

Let Y_j for $j = 1, \ldots, N$ denote the grouped counts. Let $X_j = (1, X_{1,j}, \ldots, X_{10,j})^\top$ denote the intercept term, along with indicator variables for marriage duration, place of residence, and education level of each group. Let n_j for $j = 1, \ldots, N$ denote the number of women in each group.

1. Write down the likelihood for a Poisson generalized linear model (glm) of the grouped counts assuming counts are independent and $E(Y_j) = n_j\lambda_j = n_j X_j\beta$ for a vector of unknown coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_{10})^\top$.

2. Besides a Poisson regression model, provide at least one other glm that may be appropriate for the CEB data. What are the potential pros and cons of using the Poisson glm versus your alternative glm?

The next step is to fit the Poisson glm and assess model fit. Find the fitted Poisson glm below:

```
##
## Call:
## glm(formula = round(y) ~ dur + res + educ,
##     family = poisson(link = "log"),
##     data = ceb, offset = log(n))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2960  -0.6641   0.0725   0.6336   3.6782
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05754    0.04803   1.198    0.231
## dur10-14     1.36940    0.05107  26.815  < 2e-16 ***
## dur15-19     1.61376    0.05119  31.522  < 2e-16 ***
## dur20-24     1.78491    0.05121  34.852  < 2e-16 ***
## dur25-29     1.97641    0.05003  39.501  < 2e-16 ***
## dur5-9       0.99693    0.05274  18.902  < 2e-16 ***
## resSuva     -0.15166    0.02833  -5.353 8.63e-08 ***
## resurban    -0.03924    0.02463  -1.594    0.111
## educnone    -0.02297    0.02266  -1.014    0.311
## educsec+    -0.33312    0.05390  -6.180 6.41e-10 ***
## educupper   -0.12425    0.03000  -4.142 3.44e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3731.852  on 69  degrees of freedom
## Residual deviance:   70.665  on 59  degrees of freedom
## AIC: 522.14
##
## Number of Fisher Scoring iterations: 4
```

3. Given the `R` output above, compute the relevant deviance statistic for testing whether the fitted model fits substantially better than the null model without covariates.

4. Under the null hypothesis $H_0 : \beta_1 = \cdots = \beta_{10} = 0$ what is the distribution of the above test statistic?

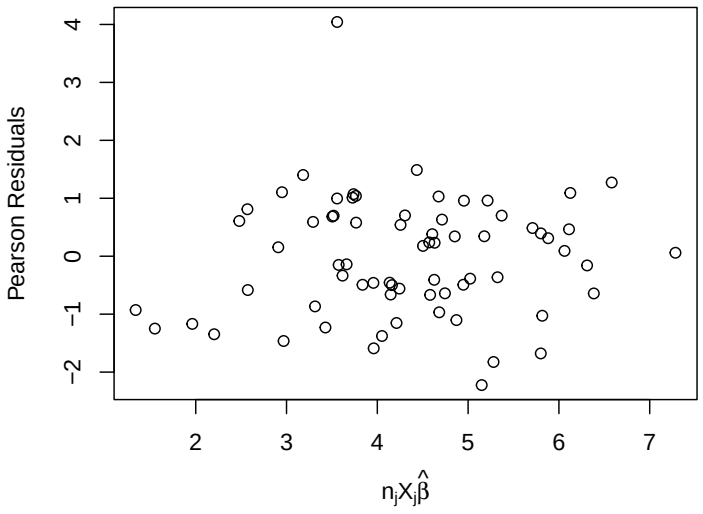

Figure 1: Pearson residuals.

In addition to model fit, we should assess the appropriateness of our chosen mean-variance relationship implied by the Poisson model.

5. What is the value of the dispersion parameter ϕ in a Poisson glm?

6. For the CEB data, Pearson and Fletcher estimates for ϕ equal about 1.21 and 1.19. Do these values tend to support the appropriateness of the Poisson model or not?

7. Examine the Pearson residual plot in Figure 1. Does it tend to support the appropriateness of the Poisson model or not?

You may have noticed one Pearson residual displayed in Figure 1 appears extreme.

8. In general, should outlier observations be removed from data analysis when using glms? Why or why not?

9. According Figure 2, should any observations be removed from the analysis?

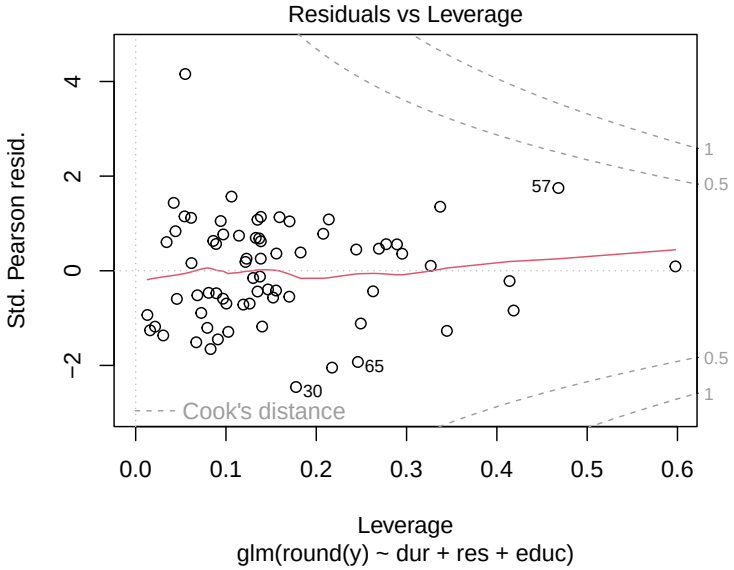

Figure 2: Residual vs. leverage plot.

Part II For each age (3 years, 5 years, 10 years, 15 years, 20 years and 25 years), the heights (in feet) of 14 Loblolly pine trees are recorded. The observed trees represent a random sample of trees rather than a set of trees of particular interest. Therefore, a model of the height-age relationship should contain random—rather than fixed—tree effects.

10. Provide the model equation for a normal linear mixed effects model of height versus age with random intercepts for trees. Define all model parameters including distributional assumptions.

11. Provide the ANOVA F test for testing the hypothesis that the random tree effects are ignorable.

12. Figure 3 below displays the observed (age, height) pairs for each tree. Different trees are identified by a seed number between 301 and 331. Based on this plot, does the model you gave in 10 seem appropriate? Why or why not?

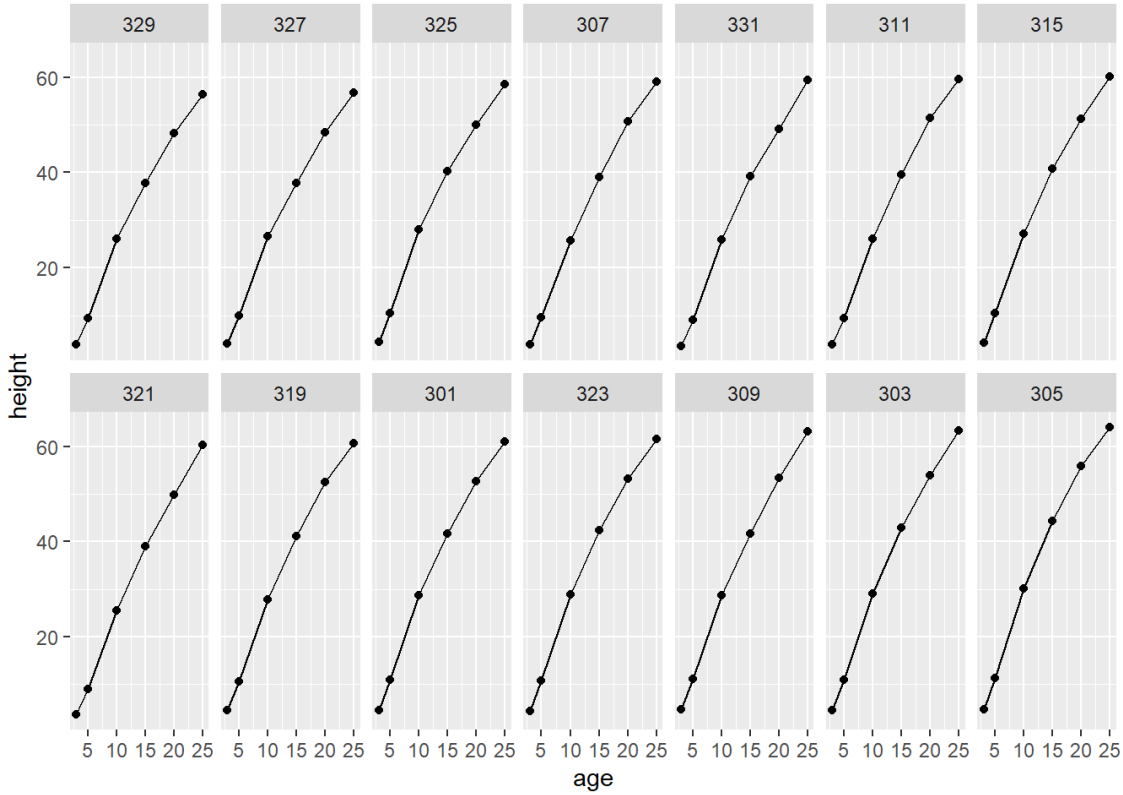

Figure 3: Height vs. age of Loblolly trees.

Suppose that after considerable work you decide to fit the following model:

$$Y = X\beta + Z\alpha + \epsilon. \tag{1}$$

Y is the 84×1 vector of tree heights. X is the 84×3 design matrix of fixed effects consisting of an intercept term along with orthogonal linear and quadratic age terms. β is the 3×1 vector of fixed effects. Z is an 84×42 design matrix of random effects. Z has a block-diagonal structure $Z = diag(Z_1, \ldots, Z_{14})$ consisting of 14 copies of a single 6×3 matrix containing a vector of ones for an intercept term and two columns for orthogonal linear and quadratic age effects. α is the 42×1 vector of random effects. And, ϵ is the 84×1 vector of random residuals. Assume $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$ independent of α. Also, assume $\alpha = (\alpha_1, \ldots, \alpha_{14})^\top$ where $\alpha_i \overset{iid}{\sim} MVN(0, \psi)$ where ψ is an unknown 3×3 positive-definite covariance matrix.

13. Provide an expression of the loglikelihood of the model in (1) as a function of the parameters (β, ψ, σ^2).

14. Suppose that (ψ, σ^2) are known. Show that the MLE of β has the form of a weighted least squares estimator $(X^\top W^{-1} X)^{-1} X^\top W^{-1} Y$ for some matrix W, and identify W.

15. Figure 4 below displays the marginal residuals $Y_{ij} - x_{ij}^\top \hat{\beta}$ by age where i and j index tree and age. Based on the plot, does the model fit the data well or not? If not, how should the model be changed?

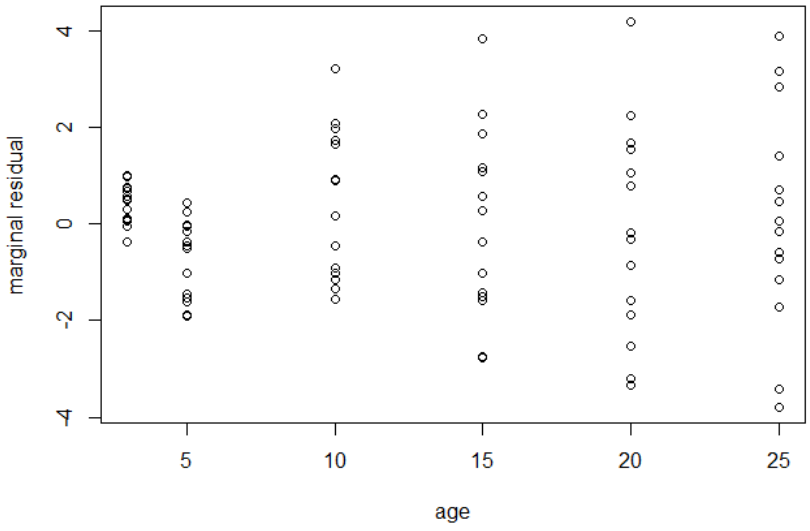

Figure 4: Maginal residuals.

16. Consider the following three models:

a) Model 1 is the quadratic mixed effects model given in (1) above.

b) Model 2 adds a cubic fixed effect to model 1.

c) Model 3 adds a cubic fixed effect and removes the quadratic random effect from model 1.

Interpret the two model comparisons given in the following R output by answering the following question. Which model would you choose and why?

```
Data: Loblolly
Models:
model1: height ~ poly(age, 2) + (poly(age, 2) | Seed)
model2: height ~ poly(age, 3) + (poly(age, 2) | Seed)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model1   12 216.71 245.88 -96.355   192.71
model2   13 175.16 206.76 -74.581   149.16 43.547  1  4.139e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Data: Loblolly
Models:
model3: height ~ poly(age, 3) + (poly(age, 1) | Seed)
model2: height ~ poly(age, 3) + (poly(age, 2) | Seed)
       npar    AIC    BIC  logLik deviance  Chisq Df Pr(>Chisq)
model3   10 211.04 235.35 -95.520   191.04
model2   13 175.16 206.76 -74.581   149.16 41.877  3  4.261e-09 ***
```

Part III The Rice dataset records the sizes of two genotypes of Asian rice (Oryza sativa) plants over time as observed using top-down, two-dimensional photography in a controlled greenhouse environment. Ten plants (five of each genotype) are observed over four equally-spaced time points for a total of 40 observations.

Consider a linear model

$$Y = X\beta + \epsilon$$

where Y is the 40×1 vector of responses (plant sizes), X is the 40×2 design matrix with a column of ones for an intercept and a column of ones and zeros indicating genotype, β is the 2×1 vector of coefficients, and ϵ is multivariate normal with mean zero.

17. Let $i = 1, \ldots, 10$ index plants and $j = 1, \ldots, 4$ index time points. For the Rice dataset, assuming linearity and normality of errors are reasonable assumptions, which linear model error structure seems more appropriate:

$$\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \qquad \text{(Gauss-Markov)}$$

or

$$(\epsilon_{i1}, \ldots, \epsilon_{i4})^\top \overset{iid}{\sim} N_4(0, \sigma^2 V) \qquad \text{(Aitken)}$$

for a positive definite matrix V? In other words, is a Gauss-Markov model or an Aitken model more appropriate. And, what would be a reasonable choice of V?

18. What is the specific advantage of using the Aitken model rather than the Gauss-Markov model?

19. Suppose you fit an Aitken model to the data for an unknown positive definite matrix

$$V = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & 1 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & 1 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & 1 \end{bmatrix}$$

by the method of maximum likelihood. State an appropriate null and alternative hypothesis for testing whether the Aitken model is more appropriate than the Gauss-Markov model. Derive a test statistic and specify its limiting distribution under the null. You are not required to derive the explicit MLEs.

Figure 5 depicts the Rice data responses over time for both genotypes.

20. With respect to the Aitken linear model, which assumptions appear to be violated?

21. Suggest a data analysis strategy to avoid the violations of these assumptions. This may include using a different model for the data.

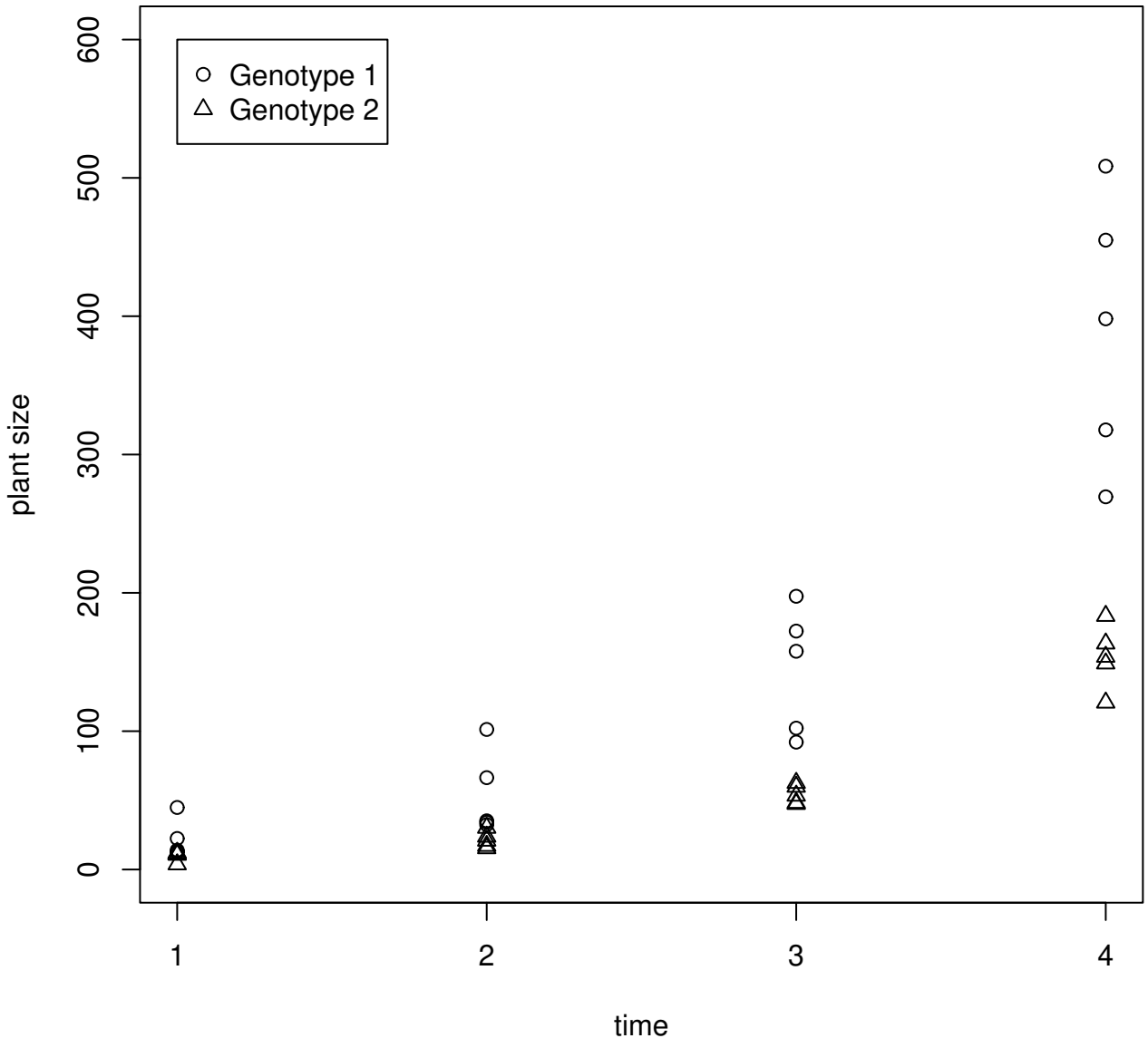

Figure 5: Rice data responses over time for both genotypes.

1. The likelihood is given by

$$L(\beta; y, x) = \prod_{j=1}^{N} (y_j!)^{-1} (n_j x_j^\top \beta)^{y_j} e^{-n_j x_j^\top \beta}$$

Some students may reflexively write down the loglikelihood, which is fine. The loglikelihood is

$$\ell(\beta; y, x) = \sum_{j=1}^{N} \left\{ -\log(y_j!) + y_j \log(n_j x_j^\top \beta) - n_j x_j^\top \beta \right\}.$$

2. Alternative glms for count data include the negative binomial and quasi-likelihood. The negative binomial can handle over-dispersed count data while quasi-likelihood, with its variable scale parameter, may handle both under- and over-dispersed data. The advantages of the Poisson include that it has one fewer parameter, and since its scale parameter equals one, likelihood ratio tests for model comparisons are more reliable than for models with an estimated scale parameter.

3. The deviance difference is $3731.852 - 70.665 = 3661.187$.

4. The deviance difference is approximately distributed as a Chi-squared random variable with 10 degrees of freedom under the null hypothesis that none of the covariates are significant.

5. 1.

6. Yes, because these are close to 1.

7. The residual plot in figure 1 reveals no change in variability versus the mean, which suggests we have adequately modeled the mean-variance relationship in the data.

8. No, not necessarily. Outliers that have high leverage (high influence observations) are concerning. Removing these from the analysis will change the results. Results will not change much if one includes versus excludes outliers without high leverage.

9. According to figure 2, none of the outliers have high leverage, and none of the observations have high influence (Cook's distance), so there is no apparent reason to consider removing any observations from the analysis.

10. The model may be written, for example,

$$Y_{ij} = \beta_0 + \beta_1 x_j + \tau_i + \epsilon_{ij}$$

where Y_{ij} is the height of tree i at age x_j; β_0 and β_1 are fixed intercept and slope effects of age on height; $\tau_i \overset{iid}{\sim} N(0, \sigma_t^2)$ is a random tree-specific intercept; and, $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ is a random residual.

11. Average over trees to obtain the following model

$$\overline{Y}_{\cdot j} = \beta_0 + \beta_1 x_j + e_j,$$

where $e_j \overset{iid}{\sim} N(0, \sigma^2/14 + \sigma_t^2)$. Compute the MSE of the full model and this aggregated model. Then,

$$F = \frac{14 \times MSE_{agg}}{MSE_{full}}$$

has and F distribution with 3 and 80 degrees of freedom under the null hypothesis $H_0 : \sigma_t^2 = 0$.

12. The model does not seem appropriate because it specifies a linear relationship between age and height while the data displays a polynomial (non-linear) relationship. We may also consider more random effects in addition to a random intercept. Even with polynomial fixed effects, a random intercept only allows up/down shifts of the polynomial over trees, but we may need more flexibility, e.g., linear and even quadratic random effects.

13. The sampling model may be written

$$Y_i \overset{ind}{\sim} MVN(x_i^\top \beta, Z_i \psi Z_i^\top + \sigma^2 I_6).$$

Therefore, the loglikelihood is

$$-\tfrac{1}{2} \log |Z\Psi Z^\top + \sigma^2 I_{84}| - \sum_{i=1}^{14} (Y_i - x_i^\top \beta)^\top (Z_i \psi Z_i^\top + \sigma^2 I_6)^{-1} (Y_i - x_i^\top \beta)$$

where $\Psi = diag(\psi, \ldots, \psi)$ is a 42×42 block diagonal matrix consisting of 14 repeated blocks of the 3×3 matrix ψ.

Equivalently, since the determinant of a block diagonal matrix is the product of determinants of the blocks, the loglikelihood may be written

$$-\tfrac{1}{2} 14 \log |Z_1 \psi Z_1^\top + \sigma^2 I_6| - \sum_{i=1}^{14} (Y_i - x_i^\top \beta)^\top (Z_i \psi Z_i^\top + \sigma^2 I_6)^{-1} (Y_i - x_i^\top \beta)$$

14. Following the likelihood calculation above, we see that the part of the loglikelihood depending on β is the quadratic form

$$(Y - X\beta)^\top (Z\Psi Z^\top + \sigma^2 I_{84})^{-1} (Y - X\beta)$$

with gradient

$$-2X^\top (Z\Psi Z^\top + \sigma^2 I_{84})^{-1} (Y - X\beta).$$

Setting to zero, we have the following normal equations

$$X^\top (Z\Psi Z^\top + \sigma^2 I_{84})^{-1} X\beta = X^\top (Z\Psi Z^\top + \sigma^2 I_{84})^{-1} Y$$

Therefore,

$$\hat{\beta} = (X^\top (Z\Psi Z^\top + \sigma^2 I_{84})^{-1} X)^{-1} X^\top (Z\Psi Z^\top + \sigma^2 I_{84})^{-1} Y.$$

the weight matrix identified with W is the covariance matrix $Z\Psi Z^\top + \sigma^2 I_{84}$.

15. There are distinct patterns in residuals as a function of age. One pattern suggests different age groups of residuals are biased up or down compared to zero. This is evidence of a poor model of the mean, and suggests including higher-order fixed effects, like a cubic term. Additionally, there is a pattern of increased variability in age, but as these are not standardized residuals, they cannot be used to assess the fit of our covariance model.

16. Both models include a likelihood ratio test with test statistic equal to twice the difference in loglikelihood of two fitted models. Under the null hypothesis the test statistic has an approximate Chi-squared distribution. These likelihood ratio tests (and corresponding null distributions) are justified by sufficiently large sample sizes, nested models, and regularity conditions. The comparison of model 1 to model 2 is a comparison of nested models involving only different fixed effects. The null hypothesis corresponds to fixing one fixed effect at zero, which is a value interior to the parameter space. The test is appropriate and the test suggests the more complex model including a cubic fixed effect fits better than the simpler quadratic model. The second comparison between models 2 and 3 corresponds to setting some covariance parameters (including one variance parameter) equal to zero. This hypothesis lies on the boundary of the parameter space, so the regularity conditions are violated and the Chi-squared null distribution is not justified. Nevertheless, the large, consistent differences in AIC, BIC, loglikelihood, and deviance suggest the model including a quadratic random effect fits better than the model without this random effect.

It is not obvious if the loglikelihood values reported are based on REML or ML (they are actually based on ML). When comparing models with different fixed effects, comparisons should be based on ML rather than REML estimates.

17. An Aitken model with AR(1) structure, i.e.,

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

is more appropriate than the Gauss-Markov model. The point is that the error covariance should reflect the fact that observations within plant (over time) are (positively) correlated.

18. OLS and GLS Point estimators of β are both unbiased and consistent. However, since the estimated standard error of the OLS estimator is generally not consistent, inferences from the Gauss-Markov model (p-values, CIs, and PIs) will be misleading. The GLS inferences, on the other hand, are trustworthy when the model is correctly specified.

19. Write

$$V = \begin{bmatrix} 1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & 1 & \sigma_{23} & \sigma_{24} \\ \sigma_{13} & \sigma_{23} & 1 & \sigma_{34} \\ \sigma_{14} & \sigma_{24} & \sigma_{34} & 1 \end{bmatrix}.$$

The Gauss-Markov model obtains when $\sigma_{ij} = 0$ in V above. These values are in the interior of the parameter space, so a likelihood ratio test is appropriate. Let

$$L_{GM}(\beta, \sigma^2) = \prod_{i=1}^{40} \phi(y_i; x_i^\top \beta, \sigma^2)$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes the normal density function. Then, $L_{GM}(\hat{\beta}_{OLS}, \frac{n-1}{n}\hat{\sigma}^2_{MSE})$ is the maximum of the likelihood under the Gauss-Markov model where $\hat{\beta}_{OLS}$ is the OLS estimator and $\hat{\sigma}^2_{MSE}$ is the mean-squared error. Let

$$L_{GLS}(\beta, \sigma^2, V) = \prod_{i=1}^{10} (2\pi)^{-4/2} det(\sigma^2 V)^{-1/2} \exp\left[-\tfrac{1}{2}(y_i - x_i^\top \beta)^\top (\sigma^2 V)^{-1}(y_i - x_i^\top \beta)\right].$$

And, let $L_{GLS}(\hat{\beta}, \hat{\sigma}^2, \hat{V})$ be the maximum of the likelihood achieved at $(\hat{\beta}, \hat{\sigma}^2, \hat{V})$. Then,

$$\Lambda = -2\log(L_{GM}(\hat{\beta}_{OLS}, \tfrac{n-1}{n}\hat{\sigma}^2_{MSE})/L_{GLS}(\hat{\beta}, \hat{\sigma}^2, \hat{V}))$$

is the likelihood ratio test statistic, which approximately follows a Chi-squared distribution with six degrees of freedom under the null hypothesis that the Gauss-Markov model is correctly specified.

20. The mean function does not appear to be linear and variance of responses tends to increase with time (heteroskedasticity).

21. Figure 5 strongly suggests a log transformation of the responses to address both of the violations.

Note that a nonlinear mean model, e.g., using a quadratic mean function in time, will not ameliorate the problem of heteroskedastic error variances. One could fit a general linear model with a different variance for each plant, but this would increase the number of model parameters by nine. Trying a log (or other Box-Cox) transformation seems like the best strategy to try next.

There are many problems that can be analyzed through the use of networks or random graph models. Recall that a graph consists of a set of nodes and a list of edges between pairs of those nodes. The nodes often represent entities such as people, cities, and so forth, and edges represent some type of relation between those nodes. In a random graph model, we often take the set of nodes to be given and fixed, and then formulate a probability model for the realization or not of potential edges that represent reciprocal relations. In your coursework you have discussed several basic structures for these types of network models, including Erdős-Rényi graph models, block or covariate models, exponential random graph models, and local structure graph models.

A basic random graph model such as those just mentioned can be extended in several ways. One extension would be to consider two graphs, formulated for different types of nodes and edges. As just one example, consider the nodes of one graph to consist of organizations, and the nodes of another graph to consist of events. Edges joining nodes in the first graph might represent some type of a relation between the organizations, and edges in the second graph might represent some underlying common characteristic of the events. Edges between nodes of the first graph and nodes of the second graph then would represent involvement of the organizations of the first graph in the events of the second graph.

Part I We first consider a two-graph model with fixed sets of nodes in each graph and undirected edges. A general structure for this scenario is as follows.

- Let nodes of the first graph be denoted as $\{u_i : i = 1, \ldots, n_1\}$ and nodes of the second graph be denoted as $\{v_i : i = 1, \ldots, n_2\}$. Assume that $n_1 \leq n_2$.
- Define random variables to represent the realization of edges in the first graph as, for $i, j = 1, \ldots, n_1$ and $i \neq j$

$$Y_{i,j} = \begin{cases} 1 & \text{if nodes } u_i \text{ and } u_j \text{ are joined by an edge,} \\ 0 & \text{otherwise.} \end{cases}$$

  There are $n_1(n_1 - 1)/2$ of these random variables.

- Define random variables to represent the realization of edges in the second graph as, for $h, \ell = 1, \ldots, n_2$ and $h \neq \ell$

$$Z_{h,\ell} = \begin{cases} 1 & \text{if nodes } v_h \text{ and } v_\ell \text{ are joined by an edge,} \\ 0 & \text{otherwise.} \end{cases}$$

  There are $n_2(n_2 - 1)/2$ of these random variables.

- Define random variables to represent the realization of edges between graphs as, for $i = 1, \ldots, n_1$ and $h = 1, \ldots, n_2$,

$$W_{i,h} = \begin{cases} 1 & \text{if nodes } u_i \text{ and } v_h \text{ are joined by an edge,} \\ 0 & \text{otherwise.} \end{cases}$$

  There are $n_1 n_2$ of these random variables.

Define the following sets:

$$\begin{aligned}\mathcal{G}_{1,1} &= \{(i,j) : i,j = 1,\ldots,n_1; i<j\}\\ \mathcal{G}_{2,2} &= \{(h,\ell) : h,l = 1,\ldots,n_2; h<l\}\\ \mathcal{G}_{1,2} &= \{(i,h) : i = 1,\ldots,n_1;\ h = 1,\ldots,n_2\}\end{aligned}$$

Note that $\mathcal{G}_{1,1}$ is the set of unique node index pairs for graph 1, $\mathcal{G}_{2,2}$ is the set of unique node index pairs for graph 2, and $\mathcal{G}_{1,2}$ is the set of unique node index pairs where one index comes from graph 1 and the other from graph 2.

1. Perhaps the simplest model that is at all realistic takes the random variables $Y_{i,j}$, $Z_{h,\ell}$ and $W_{i,h}$ to be independent for all values of i, j, h and ℓ, and specifies that,

$$\begin{aligned}Pr(Y_{i,j}=1) &= p;\ \ (i,j)\in\mathcal{G}_{1,1},\\ Pr(Z_{h,\ell}=1) &= q;\ \ (h,\ell)\in\mathcal{G}_{2,2},\\ Pr(W_{i,h}=1) &= d;\ \ (i,h)\in\mathcal{G}_{1,2}.\end{aligned} \tag{1}$$

Find the maximum likelihood estimators of p, q, and d for this model.

2. Suppose we would like to conduct a simulation-based assessment of the model in Question 1. Define the test quantities,

$$\begin{aligned}T_{1,1} &= \sum_{(i,j)\in\mathcal{G}_{1,1}} y_{i,j},\\ T_{2,2} &= \sum_{(h,\ell)\in\mathcal{G}_{2,2}} z_{h,\ell},\\ T_{1,2} &= \sum_{(i,h)\in\mathcal{G}_{1,2}} w_{i,h},\end{aligned}$$

Why would these test quantities **not** lead to a meaningful model assessment?

3. Suppose that an alternative to model (1) is to recognize that there are k_1 blocks of nodes in graph 1, and there are k_2 blocks of nodes in graph 2, such that nodes within a block share a common characteristic (e.g., geographic region) that is not shared by nodes in other blocks. One model this could lead to is to define edges within blocks and between blocks within each graph, and then also between blocks of different graphs. To simplify, consider a model with k_1 distinct probabilities for edge realization between nodes within the k_1 blocks of graph 1, k_2 distinct probabilities for edge realization between nodes within the k_2 blocks of graph 2, only one probability for edge realization between nodes of different blocks of graph 1, and similarly for graph 2. Add to these one additional probability that applies to the realization of any potential edge between a node in graph 1 with a node in graph 2. This model would contain $k_1 + k_2 + 3$ distinct probabilities for edge realizations. With an eye toward this alternative model, suggest a test quantity for use in a simulation-based assessment of model (1) that would result in a meaningful procedure.

Part II Any number of potential applications of two-graph models might involve processes that are *contagious* in the sense that nodes with high degree (number of edges they are involved in) tend to be connected with other nodes of high degree. One possible model to reflect this structure can be formulated as follows. Continue to assume undirected edges unless directed edges are explicitly specified.

We will use the same random variables $Y_{i,j}$, $Z_{h,\ell}$ and $W_{i,h}$ and index sets $\mathcal{G}_{1,1}$, $\mathcal{G}_{2,2}$ and $\mathcal{G}_{1,2}$ defined previously. Let $p(x)$ denote generic notation for a probability mass function for a random variable X. Assign $Y_{i,j}$, $Z_{h,\ell}$ and $W_{i,h}$ the distributions, for $0 < \theta_{i,j} < 1$, $0 < \eta_{h,\ell} < 1$, and $0 < \lambda_{i,h} < 1$,

$$\begin{aligned} p_{i,j}(y|\theta_{i,j}) &= \theta_{i,j}^{y}(1-\theta_{i,j})^{1-y}; \quad y = 0, 1; \quad (i,j) \in \mathcal{G}_{1,1} \\ p_{h,\ell}(z|\eta_{h,\ell}) &= \eta_{h,\ell}^{z}(1-\eta_{h,\ell})^{1-z}; \quad z = 0, 1; \quad (h,\ell) \in \mathcal{G}_{2,2} \\ p_{i,h}(w|\lambda_{i,h}) &= \lambda_{i,h}^{w}(1-\lambda_{i,h})^{1-w}; \quad w = 0, 1; \quad (i,h) \in \mathcal{G}_{1,2} \end{aligned} \tag{2}$$

where,

$$\begin{aligned} \log\left(\frac{\theta_{i,j}}{1-\theta_{i,j}}\right) &= \kappa_y + \beta\left(\sum_{k\neq i,j} y_{i,k} + \sum_{k\neq i,j} y_{k,j}\right) \\ \log\left(\frac{\eta_{h,\ell}}{1-\eta_{h,\ell}}\right) &= \kappa_z + \alpha\left(\sum_{k\neq h,\ell} z_{h,k} + \sum_{k\neq h,\ell} z_{k,\ell}\right) \\ \log\left(\frac{\lambda_{i,h}}{1-\lambda_{i,h}}\right) &= \kappa_w + \gamma\left(\sum_{k\neq i,h} w_{i,k} + \gamma_2 \sum_{k\neq i,h} w_{k,h}\right). \end{aligned} \tag{3}$$

4. The structure of the model in (2) and (3) is similar to a set of three logistic regressions or generalized linear models (glms) with binary random components and logit link functions. In words, describe what the "covariates" inside parentheses in (3) represent.

5. Briefly explain why the Fisher Scoring algorithm typically used to estimate basic glms would not produce maximum likelihood estimates of the parameters in (3).

6. If we would use that algorithm anyway, what is the objective function being maximized? Is there any justification we could give for approaching estimation with that procedure?

7. If one were to attempt to invoke a result of asymptotic normality for estimators of κ_y, κ_z, κ_w, α, β, and γ. How would you present the asymptotic context? In what way might this impact the approach you would take to compute interval estimates of these parameters?

8. After estimation of the the model in (2) and (3), what would you examine or construct to determine if the estimated model has resulted in a finding that there are dependencies among the random variables involved? Justify your answer.

9. It is claimed that the usual deviance used in basic generalized linear models would be a suitable quantity as the basis for assessment of this model. Give a brief argument that either supports or contradicts this claim.

10. Outline an algorithm for simulation of data from this model, assuming we have values (perhaps estimated) for the parameters.

Part III Now suppose we have only one graph, with n nodes. Consider a scenario in which the edges of our graph are directed, but each pair of nodes can potentially have edges in each direction. So there is a potential edge from node u_i to node u_j and another potential edge from node u_j to node u_i. For context, suppose nodes are now countries and edges represent exports of goods from a source country to a receiving country. There are any number of notational systems that could be used in this situation. We will use the following one. Define random variables that correspond to potential edges from node u_i to u_j as, for $i, j = 1, \ldots, n;\ i \neq j$,

$$Y_{i,j} = \begin{cases} 1 & \text{if there is an edge from } u_i \text{ to } u_j, \\ 0 & \text{otherwise.} \end{cases}$$

Assume that the probability mass functions of the $Y_{i,j}$ are given as,

$$p_y(y|\theta_{i,j}) = \theta_{i,j}^y (1-\theta_{i,j})^{1-y};\quad y = 0, 1;\ \ i, j = 1, \ldots, n;\ i \neq j.$$

We again wish to use the concept of contagion in further modeling of the parameters $\theta_{i,j}$. It is believed that the number of edges emanating from u_i might have one effect on the probability of an edge from u_i to u_j, while the number of edges terminating in u_j might have a different effect.

11. Modify the form of (3) to complete a model for this situation. Note that it is only the first line of (3) that is of concern.

12. Now consider a situation in which there can be only one edge between u_i and u_j but this edge is labeled. For example, if the nodes again represent organizations, edges may be either cooperative (type 1) or antagonistic (type 2) in nature. Here, there cannot be both a type 1 edge and a type 2 edge between nodes u_i and u_j. This is similar to a graph with directed edges, but for which there can be only one edge between any two nodes. Let $\mathcal{G}$ denote the set of all $n(n-1)/2$ unique pairs of node indices (i, j). To develop a model for this situation we could define random variables as, for $(i, j) \in \mathcal{G}$,

$$Y_{i,j} = \begin{cases} 1 & \text{if there is an edge of type 1 between } u_i \text{ and } u_j, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$Z_{i,j} = \begin{cases} 1 & \text{if there is an edge of type 2 between } u_i \text{ and } u_j, \\ 0 & \text{otherwise.} \end{cases}$$

For this situation, briefly explain what would be the primary challenge for model formulation.
Hint: The marginal support of the distributions of each $Y_{i,j}$ and each $Z_{i,j}$ is $\Omega_{i,j} = \{0, 1\}$. Consider the relation between these marginal supports and the support of what the joint distribution of all $Y_{i,j}$ and $Z_{i,j}$ must be.

Part I

1. Writing the probabilities of expression (1) in the question as probability mass functions for $Y_{i,j}$, $Z_{h,\ell}$ and $W_{i,h}$, we have,

$$\begin{aligned} p_{i,j}(y|p) &= p^y(1-p)^{1-y}; \quad y \in \{0,1\}, \\ p_{h,\ell}(z|q) &= q^z(1-q)^{1-z}; \quad w \in \{0,1\}, \\ p_{i,h}(w|d) &= d^w(1-d)^{1-w}; \quad w \in \{0,1\}. \end{aligned} \tag{1}$$

For observations $\{y_{i,j} : (i,j) \in \mathcal{G}_{1,1}\}$, $\{z_{h,\ell} : (h,\ell) \in \mathcal{G}_{2,2}\}$ and $\{z_{i,h} : (i,h) \in \mathcal{G}_{1,2}\}$, the log likelihood is,

$$\begin{aligned} \ell(p,q,d) &= \sum_{(i,j)\in\mathcal{G}_{1,1}} [y_{i,j}\log(p) + (1-y_{i,j})\log(1-p)] \\ &+ \sum_{(h,\ell)\in\mathcal{G}_{2,2}} [z_{h,\ell}\log(q) + (1-z_{h,\ell})\log(1-q)] \\ &+ \sum_{(i,h)\in\mathcal{G}_{1,2}} [w_{i,h}\log(d) + (1-w_{i,h})\log(1-d)]. \end{aligned} \tag{2}$$

This is a problem with regular exponential family distributions so maximum likelihood estimates may be determined by calculating and solving the usual score equations. To simplify notation let,

$$\begin{aligned} T_{1,1} &= \sum_{(i,j)\in\mathcal{G}_{1,1}} y_{i,j} \\ T_{2,2} &= \sum_{(h,\ell)\in\mathcal{G}_{2,2}} z_{h,\ell} \\ T_{1,2} &= \sum_{(i,h)\in\mathcal{G}_{1,2}} w_{i,h}. \end{aligned}$$

Recall from the question that $|\mathcal{G}_{1,1}| = n_1(n_1-1)/2$, $|\mathcal{G}_{2,2}| = n_2(n_2-1)/2$, and $|\mathcal{G}_{1,2}| = n_1 n_2$. Then,

$$\begin{aligned} \frac{\partial\ell(p,q,d)}{\partial p} &= \frac{T_{1,1}}{p} - \frac{n_1(n_1-1)/2 - T_{1,1}}{1-p}, \\ \frac{\partial\ell(p,q,d)}{\partial q} &= \frac{T_{2,2}}{q} - \frac{n_2(n_2-1)/2 - T_{2,2}}{1-q}, \\ \frac{\partial\ell(p,q,d)}{\partial p} &= \frac{T_{1,2}}{d} - \frac{n_1 n_2 - T_{1,2}}{1-d}, \end{aligned}$$

which leads to

$$\begin{aligned} \hat{p} &= \frac{2T_{1,1}}{n_1(n_1-1)} \\ \hat{q} &= \frac{2T_{2,2}}{n_2(n_2-1)} \\ \hat{d} &= \frac{T_{1,2}}{n_1 n_2} \end{aligned} \tag{3}$$

2. As is clear from the answer to Question 1, these are sufficient statistics. Simulated data from nearly any model, no matter how poor a representation of the data, will be able to re-create sufficient statistics. Consider, for example, fitting a one-sample normal model to a set of that that is clearly bimodal. Simulated data sets from the horribly flawed fitted normal model will still have sample means that are similar to those in the data.

3. Let $b_s : s = 1, \ldots k_1$ denote the blocks in graph 1 and let $a_r : r = 1, \ldots, k_2$ denote the blocks in graph 2. The basic model takes the probability of edge realization to be constant within a graph and constant between graphs. The block model being considered as an alternative to that basic model takes the probability of edge realization to vary among blocks for each graph, but to be constant between blocks and also between graphs. We need this difference to be reflected in a test statistic. A useful test statistic for a simulation-based model assessment could be developed as follows. Recall that nodes in graph 1 are denoted as u_i and nodes in graph 2 are denoted as v_h. Let $I(A)$ denote the identity function that assumes the value 1 if A is true and the value 0 otherwise.

Define the quantities,

$$\begin{aligned} T_s &= \sum_{(i,j)\in\mathcal{G}_{1,1}} y_{i,j} I(u_i \in b_s) I(u_j \in b_s);\ s = 1, \ldots, k_1 \\ T_r &= \sum_{(h,\ell)\in\mathcal{G}_{2,2}} z_{i,j} I(v_i \in a_r) I(v_j \in a_r);\ r = 1, \ldots, k_2 \end{aligned} \tag{4}$$

Notice that the number of edges occurring between nodes in different blocks within each graph is not really relevant to the difference between the basic model and the block model, nor is the number of edges occurring between nodes in different graphs.

Let T_s^a and T_r^a denote the values of these statistics in the actual data, and let $T_s^*(m)$ and $T_r^*(m)$ denote their values in a set of data simulated from the fitted model and indexed by m. Now, using $var(T)$ to denote the sample variance of values contained in a set T, define the statistics,

$$\begin{aligned} V_1(m) &= var(T_p^*(m)) \\ V_2(m) &= var(T_r^*(m)) \\ V_1^a &= var(T_p^a) \\ V_2^a &= var(T_r^a) \end{aligned} \tag{2}$$

You could use these statistics in several ways. One would be to consider the two types of edges separately and compute $p-$values as,

$$\begin{aligned} p_1 &= \frac{1}{M}\sum_{m=1}^{M} I\left[V_1^a \le V_1^*(m)\right] \\ p_2 &= \frac{1}{M}\sum_{m=1}^{M} I\left[V_2^a \le V_2^*(m)\right]. \end{aligned}$$

The form of these $p-$values is determined by the fact that greater heterogeneity among blocks in the proportion of realized edges results in larger variances in the statistics (5).

Another possible use would be to combine these statistics as

$$D(m) = \max\{V_1(m), V_2(m)\} \text{ and } D^a = \max\{V_1^a, V_2^a\}$$

or as,

$$D(m) = V_1(m) + V_2(m) \text{ and } D^a = V_1^a + V_2^a.$$

Whichever test statistic we choose, a simulation-based $p-$value would be computed as,

$$p = \frac{1}{M} \sum_{m=1}^{m} I\left[D^a \leq D(m)\right],$$

again, because greater heterogeneity among edge proportions among blocks is reflected in larger values of the statistics.

Part II

4. The "covariates" contained in the question represent the number of other (realized) edges that the two nodes in question are involved in. In particular, for the first line of expression (3) in the question,

$$\sum_{k \neq i,j} y_{i,k} \quad = \quad \text{the number of realized edges of graph 1 for which node } u_i \text{ is one of the terminals,}$$

$$\sum_{k \neq i,j} y_{k,j} \quad = \quad \text{the number of realized edges of graph 1 for which node } u_j \text{ is one of the terminals.}$$

and similarly for edges in graph 2 (second line) and edges between graphs (third line).

5. The usual Fisher Scoring algorithm used in basic generalized linear models is not applicable to estimation of the parameters κ_y, κ_z, κ_w, β, α, and γ because the the distributions in expression (2) of the question do not define marginal distributions, and the random variables $Y_{i,j}$, $Z_{h,\ell}$ and $W_{i,h}$ are not all independent.

6. If we would use the Fisher Scoring glm algorithm for this problem anyway, the objective function being maximized would be,

$$\begin{aligned} Q \quad = \quad & \sum_{(i,j) \in \mathcal{G}_{1,1}} y_{i,j} \log(\theta_{i,j}) + (1 - y_{i,j}) \log(1 - \theta_{i,j}) \\ + \quad & \sum_{(h,\ell) \in \mathcal{G}_{2,2}} z_{h,\ell} \log(\eta_{h,\ell}) + (1 - z_{h,\ell}) \log(1 - \eta_{h,\ell}) \\ + \quad & \sum_{(i,h) \in \mathcal{G}_{1,2}} w_{i,h} \log(\lambda_{i,h}) + (1 - w_{i,h}) \log(1 - \lambda_{i,h}), \end{aligned}$$

where,

$$\theta_{i,j} = \frac{\exp\left[\kappa_y + \beta\left(\sum_{k\neq i,j} y_{i,k} + \sum_{k\neq i,j} y_{k,j}\right)\right]}{1+\exp\left[\kappa_y + \beta\left(\sum_{k\neq i,j} y_{i,k} + \sum_{k\neq i,j} y_{k,j}\right)\right]},$$

$$\eta_{h,\ell} = \frac{\exp\left[\kappa_z + \alpha\left(\sum_{k\neq h,\ell} z_{h,k} + \sum_{k\neq h,\ell} z_{k,\ell}\right)\right]}{1+\exp\left[\kappa_z + \alpha\left(\sum_{k\neq h,\ell} z_{h,\ell} + \sum_{k\neq h,\ell} z_{k,\ell}\right)\right]},$$

$$\lambda_{i,h} = \frac{\exp\left[\kappa_w + \gamma\left(\sum_{k\neq i,h} y_{i,k} + \sum_{k\neq i,j} y_{k,h}\right)\right]}{1+\exp\left[\kappa_w + \gamma\left(\sum_{k\neq i,h} y_{i,k} + \sum_{k\neq i,h} y_{k,h}\right)\right]}.$$

There is justification for maximizing this objective function and that is due to the fact that it qualifies as a composite likelihood based on full conditional distributions, which is what the mass functions in expression (2) of the question represent. The model also qualifies as a Markov random field model in which neighborhoods of potential edges are all other potential edges in the same group – in graph 1 for the $Y_{i,j}$, in graph 2 for the $Z_{h,\ell}$, and between graphs for the $W_{i,j}$. Thus, the composite likelihood just described also qualifies as the original pseudo-likelihood of Besag.

7. The asymptotic context would almost certainly be that of a (hypothetical or superpopulation) expanding lattice, rather than a repeating lattice. It is difficult to envision a situation in which repeated realizations of the the same probabilistic structure could be obtained, particularly with graph 2 in which the nodes are events. It is unlikely the same events could occur repeatedly. The impact on interval estimation would be that computation of Godambe Information might be difficult or unstable (without replication), motivating the use of parametric bootstrap for direct computation of intervals.

8. The obvious choice is interval estimation of the parameters β, α, and γ in the model. If all of these are zero, then there are no dependencies among the random variables $Y_{i,j}$, $Z_{h,\ell}$, and $W_{i,h}$, at least dependencies of the type being represented in this model.

9. This claim can be supported. The full conditional distributions givin in expression (2) of the question are exponential dispersion family forms. As a result, deviance is a valid measure of discrepancy between a fitted model and the maximal (or saturated) model. Unless there is complete independence among all random variables involved in the model none of the results sometimes attached to deviance in glms will hold, but deviance itself is still a legitimate discrepancy measure and could be used in a simulation-based model assessment procedure.

10. Simulation of data from the model would most naturally take the form of a Gibbs Sampling algorithm, again due to the nature of the probability mass functions in the model representing full conditional distributions. Thus, an outline of a suitable algorithm would be as follows:

a) Determine starting values for all of the variables $\{y_{i,j} : (i,j) \in \mathcal{G}_{1,1}\}$, $\{z_{h,\ell} : (h,\ell) \in \mathcal{G}_{2,2}\}$, and $\{w_{i,h} : (i,h) \in \mathcal{G}_{1,2}\}$. This can be done in an arbitrary manner, setting all values to 0, setting all values to 1, or through a random sample from a binary distribution with some parameter (such as 0.50). Select a value for number of cycles to complete, M.

b) Chose an arbitrary order for the variables within each group, and an arbitrary order for the groups. This could be done once and then used repeatedly, or could be done at the start of each Gibbs cycle. For ease of presentation, suppose we will consider the Y, Z, and W groups in that order, fixed for the entire algorithm. We can re-index the variables as $Y_1, \ldots, Y_{n_1}$, where $n_1 = |\mathcal{G}_{1,1}|$, $Z_1, \ldots, Z_{n_2}$, where $n_2 = |\mathcal{G}_{2,2}|$, and $W_1, \ldots, W_{n_3}$, where $n_3 = |\mathcal{G}_{1,2}|$. The full conditional mass functions of these variables must be indexed in a matched fashion sot that Y_1 and $p_1(y)$ corresponds to the same pair of nodes (i, j), and similarly all other variables. Increase the value of m by 1.

c) For $i = 1, \ldots n_1$, simulate a value y_i^* from $pi(y|\{y_j : j \neq i\})$ and replace y_i with y_i^*.

d) For $k = 1, \ldots, n_2$, simulate a value z_k^* from $p_k(z|\{z_h : h \neq k\})$ and replace z_k with z_k^*.

e) For $r = 1, \ldots, n_3$, simulate a value w_r^* from $p_r(w|\{w_s : s \neq r\})$ and replace w_r with w_r^*.

f) Increase the value of m 1. If $m < M$ return to step 2 (sort of a random scan) or step 3 (for a fixed scan). If $m = M$ stop and take the current sets of values to be a realization from the model.

If multiple values are desired, as in a simulation-based model assessment, continue the algorithm and collect subsequent sets of values rather than stopping in step 6, or possibly "thin" by collecting only values separated by a certain number of cylces.

Part III

11. We would use the model, for $i, j = 1, \ldots, n;\ i \neq j$,

$$\log\left(\frac{\theta_{i,j}}{1-\theta_{i,j}}\right) = \kappa + \alpha\left(\sum_{k\neq i,j} y_{i,k}\right) + \beta\left(\sum_{k\neq i,j} y_{k,j}\right).$$

Here, the sum $\sum_{k\neq i,j} y_{i,k}$ represents the number of edges originating at node u_i, (other than to u_j) while the sum $\sum_{k\neq i,j} y_{k,j}$ represents the number of edges terminating at node u_j (other than from u_i).

12. Clearly the random variables $Y_{i,j}$ and $Z_{i,j}$ cannot be independent and we will need to specify some structure for the parameters of binary distributions. But the greatest challenge to model development in this situation is ensuring that a joint distribution for the entire collection of $Y_{i,j}$ and $Z_{i,j}$ exists and is compatible with whatever model structure we develop. This is because we cannot rely on the usual positivity condition that the joint support is given as the Cartesian product of the marginal supports. Another way to say this same thing is that there exist *forbidden states* in the joint support.