Note that no calculators will be used for this exam. Answers that would typically require calculation should be written to make clear the calculation that an examinee would have done if a calculator were available.

Long before the pandemic hit in 2020, some companies had had interest in exploring the option of work from home (WFH; also called telecommuting or telework) to save office space and employees' commuting time. An article published in The Quarterly Journal of Economics in 2015 presented a study to investigate the efficacy of WFH as a management practice. This problem set is based on a hypothetical study similar to the published one.

A travel agency does its business on the telephone to make reservations and obtain tickets for clients. The agency generates revenue through commissions from hotels, airlines, and tour operators. Call center representatives are organized into teams of 10~16 people, grouped by department (hotel or airline) and type of work. One type of work in each of the two departments is to take orders. The order-takers answer customer calls, take orders, and enter them into the company's information system.

Part I

This travel agency conducted a study to check the effect of WFH on the productivity of order-takers. To do this, they first identified employees who were qualified to take part in the study by virtue of having broadband internet access and a private room at home in which they could work. Half of these identified employees were randomly selected to work from home (WFH), and the other half stayed in the office to act as the control group (CON). Office and home workers used the same IT equipment, faced the same work order flow from a common central server, carried out the same tasks, and were compensated under the same pay system. Hence, the only difference between the two groups was the location of work.

We look at a subset of the data generated by order-takes from the airline department. Order-takers were managed in teams with each team having its own team leader and the same work schedule. Each team had multiple employees in both treatment groups (WFH and CON). For this exam, we use a dataset representing 20 teams randomly sampled from the airline department. For each of the 20 teams, we randomly picked an employee from the WFH group and randomly picked another employee from the CON group. One of order-takers' key performance measures was the number of phone calls answered. The study lasted several months, and the manager was interested in checking the effect of WFH on the average number of phone calls per week for order-takers. Here are some summary statistics:

Table 1: Summary Statistics for Number of Phone Calls per Week for 20 Teams of Order-Takers

Treatment	Sample Mean	Sample Standard Deviation
WFH	400	80
CON	360	70

The sample correlation coefficient between weekly phone call numbers for WFH employee and CON employee in the same team is 0.12.

1. Why is this study an experiment and not an observational study?
2. Identify the experimental units and the experimental design of this study.
3. To check whether WFH affects the mean number of phone calls per week for order-takers, you are asked to conduct a *t*-test for the null hypothesis $\mu_{WFH} = \mu_{CON}$ versus the alternative hypothesis $\mu_{WFH} \neq \mu_{CON}$, where μ_{WFH} and μ_{CON} denote the mean number of phone calls per week for order-takers in the WFH and CON groups, respectively. Conduct an appropriate *t*-test and report the following:
   - **(a)** The point estimate for the difference in the mean numbers of phone calls per week for order-takers between WFH and CON.
   - **(b)** The standard deviation for the point estimate in **3(a)**.
   - **(c)** The *t*-statistic.
   - **(d)** The degrees of freedom.
4. List the assumptions for the *t*-test you conducted in **Problem 3**.
5. For each assumption you listed in **Problem 4,** check whether it is appropriate based on the study design, or state how you could check it if you were given all 40 observations.
6. Suppose that you are asked to use ANOVA and conduct an appropriate *F*-test to test the same hypotheses specified in **Problem 3**. Fill in the degrees of freedom (DF) of the ANOVA Table below.

Source	**DF**
Team	
Treatment	
Error	
Total	

7. Let Y_{ij} be the observed number of phone calls per week for the employee in the *i*-th treatment group (*i* = WFH or CON) and *j*-th team. Write the linear model for Y_{ij} corresponding to the ANOVA table in **Problem 6**, and define the terms in you model.
8. For each term in the model you wrote in **Problem 7**, identify whether it is a fixed effect or a random effect. Justify your answer.
9. Conduct an appropriate *F*-test for the same hypotheses specified in **Problem 3** and report the following:
   - **(a)** The *F*-statistic.
   - **(b)** The degrees of freedom.
10. What are the assumptions for the linear model under which your *F*-test in **Problem 9** is valid?
11. Are the assumptions for your *t*-test in **Problem 3** and the assumptions for your *F*-test in **Problem 9** the same or not? If they are not the same, which one is more general (makes fewer assumptions)?

Part II

The travel agency was also interested in how employees' well-being was affected by WFH. A standard employee satisfaction survey was given to all 200 employees who participated in this study. The general satisfaction measures were summarized in a score scaled to 1-100. Other than the treatment (WFH or CON), values for several other variables were also observed because of their potential effect on work satisfaction. These were gender, age, marital status, having children or not, and commuting time to work (in minutes) among others. Some pre-analysis has shown that neither team nor department (hotel or airline) affected the satisfaction score, so these two variables were not included for the analysis of satisfaction scores. We work on a subset of 20 employees randomly picked from all participants of this study, and a partial dataset is given in Table 2. The complete dataset is given in the SAS code on **page 5**.

Table 2: Self-reported Work Satisfaction Data

Employee	Satisfaction Score	Treatment	Gender	Age	Marital Status	Children	Commuting Time
1	61	WFH	Male	20	Single	No	50
2	65	CON	Female	25	Single	No	90
…	…		…	…	…	…	…
20	69	CON	Male	28	Married	Yes	100

Now, let Y_k be the satisfaction score for the *k*-th employee where *k=1, …, 20*. And let

$$x_{1k} = \begin{cases} 1 & \text{if the } k\text{-th employee was in the WFH group} \\ 0 & \text{if the } k\text{-th employee was in the CON group,} \end{cases}$$

$$x_{2k} = \begin{cases} 1 & \text{if the } k\text{-th employee was a female} \\ 0 & \text{if the } k\text{-th employee was a male,} \end{cases}$$

x_{3k} = age (in years) of the *k*-th employee,

$$x_{4k} = \begin{cases} 1 & \text{if the } k\text{-th employee was married} \\ 0 & \text{if the } k\text{-th employee was NOT married,} \end{cases}$$

$$x_{5k} = \begin{cases} 1 & \text{if the } k\text{-th employee had children} \\ 0 & \text{if the } k\text{-th employee did not have children,} \end{cases}$$

and, $x_{6k} = \begin{cases} 1 & \text{if the } k\text{-th employee has} \geq 90 \text{ minutes commuting time} \\ 0 & \text{if the } k\text{-th employee has} < 90 \text{ minutes commuting time.} \end{cases}$

The following regression models were fit to the dataset:

$$Y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k} + \varepsilon_k \quad \text{Model (1)}$$

$$Y_k = \beta_0 + \beta_1 x_{1k} + \varepsilon_k \quad \text{Model (2)}$$

$$\sqrt{Y_k} = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \beta_4 x_{4k} + \beta_5 x_{5k} + \beta_6 x_{6k} + \varepsilon_k \quad \text{Model (3)}$$

$$\sqrt{Y_k} = \beta_0 + \beta_1 x_{1k} + \varepsilon_k \quad \text{Model (4)}$$

where all the regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5,$ and β_6 are unknown real-valued parameters, and the ε_k are iiid $N(0, \sigma^2)$ random variables.

SAS code for analyzing this dataset is on **page 5.** Partial outputs from SAS are on **pages 6-10**.

12. Give the 1st, 2nd, and 20th rows of the design matrix **X** for model (1) corresponding to employees 1, 2, and 20 (the observations listed in Table 2).

13. Interpret the estimated value of the parameter β_1 for model (1) in the context of this study.

14. Based on SAS output for model (1) on **page 6**, conduct a hypothesis test to check whether WFH improved the mean work satisfaction score or not. Report the following:
- **(a)** The null and alternative hypotheses.
- **(b)** The value of your test statistic.
- **(c)** The degrees of freedom.
- **(d)** The *p*-value.
- **(e)** Your conclusion.

15. Based on the diagnostic plots for model (1) on **page 9**, do you have any concerns about the appropriateness of the assumptions for model (1) for this dataset? Explain.

16. Given the SAS output for fitting models (1) and (2) using SAS code on **pages 6-7**, conduct an *F*-test to check whether $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$ or not. Report the following:
- **(a)** Your *F*-statistic.
- **(b)** The degrees of freedom.

17. For the 20 employees included in this subset, all children belong to married employees. If we include an interaction term between X_4 (marital status) and X_5 (having children or not), do you think the regression coefficient for the interaction term will be estimable or not? Justify your answer.

18. Based on model selection criteria and model diagnostics on **pages 8-10**, which model do you suggest using for this data analysis?

SAS code

```
data wfh;
  input y sqrty x1 x2 x3 x4 x5 x6;
  datalines;
 61 7.810250  1  0 20  0  0  0
 65 8.062258  0  1 25  0  0  1
 56 7.483315  0  1 19  0  0  0
 63 7.937254  1  0 21  1  0  0
 57 7.549834  0  1 19  0  0  0
 66 8.124038  1  1 24  0  0  0
 60 7.745967  1  1 19  0  0  0
 62 7.874008  1  1 20  0  0  0
 65 8.062258  0  1 27  0  0  0
 62 7.874008  1  1 19  0  0  0
 71 8.426150  1  1 29  0  0  0
 65 8.062258  1  0 21  1  0  0
 64 8.000000  0  0 29  0  0  0
 61 7.810250  0  0 24  0  0  0
 72 8.485281  1  1 30  0  0  0
 69 8.306624  1  1 26  0  0  0
 70 8.366600  1  1 24  0  0  1
 72 8.485281  0  1 29  1  1  1
 69 8.306624  1  1 21  1  1  1
 69 8.306624  0  0 28  1  1  1
run;

proc reg data=wfh corr plots=diagnostics;
  model y = x1-x6 ;
  title"results for Model (1)";
run;
proc reg data=wfh plots=diagnostics ;
  model y = x1;
  title"results for Model (2)";
run;

proc reg data=wfh corr plots=diagnostics;
  model sqrty = x1-x6;
  title"results for Model (3)";
run;
proc reg data=wfh  plots=diagnostics ;
  model sqrty = x1;
  title"results for Model (4)";
run;
```

Partial SAS output for fitting Model (1) in Part II

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model			70.709	105.71	<.0001
Error			0.669		
Corrected Total					

Root MSE	0.81788	**R-Square**	0.9799
Dependent Mean	64.95000	**Adj R-Sq**	0.9706
Coeff Var	1.25924		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	36.36	1.361	26.71	<.0001
x1	1	3.18	0.825		0.0002
x2	1	1.50	0.506	2.96	0.0111
x3	1	0.982	0.051	19.13	<.0001
x4	1	2.17	0.790	2.75	0.0166
x5	1	0.159	1.057	0.15	0.8830
x6	1	3.15	0.640	4.93	0.0003

Partial SAS output for fitting Model (2) in Part II

Number of Observations Read	20
Number of Observations Used	20

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model			23.40833	1.03	0.3239
Error			22.75231		
Corrected Total					

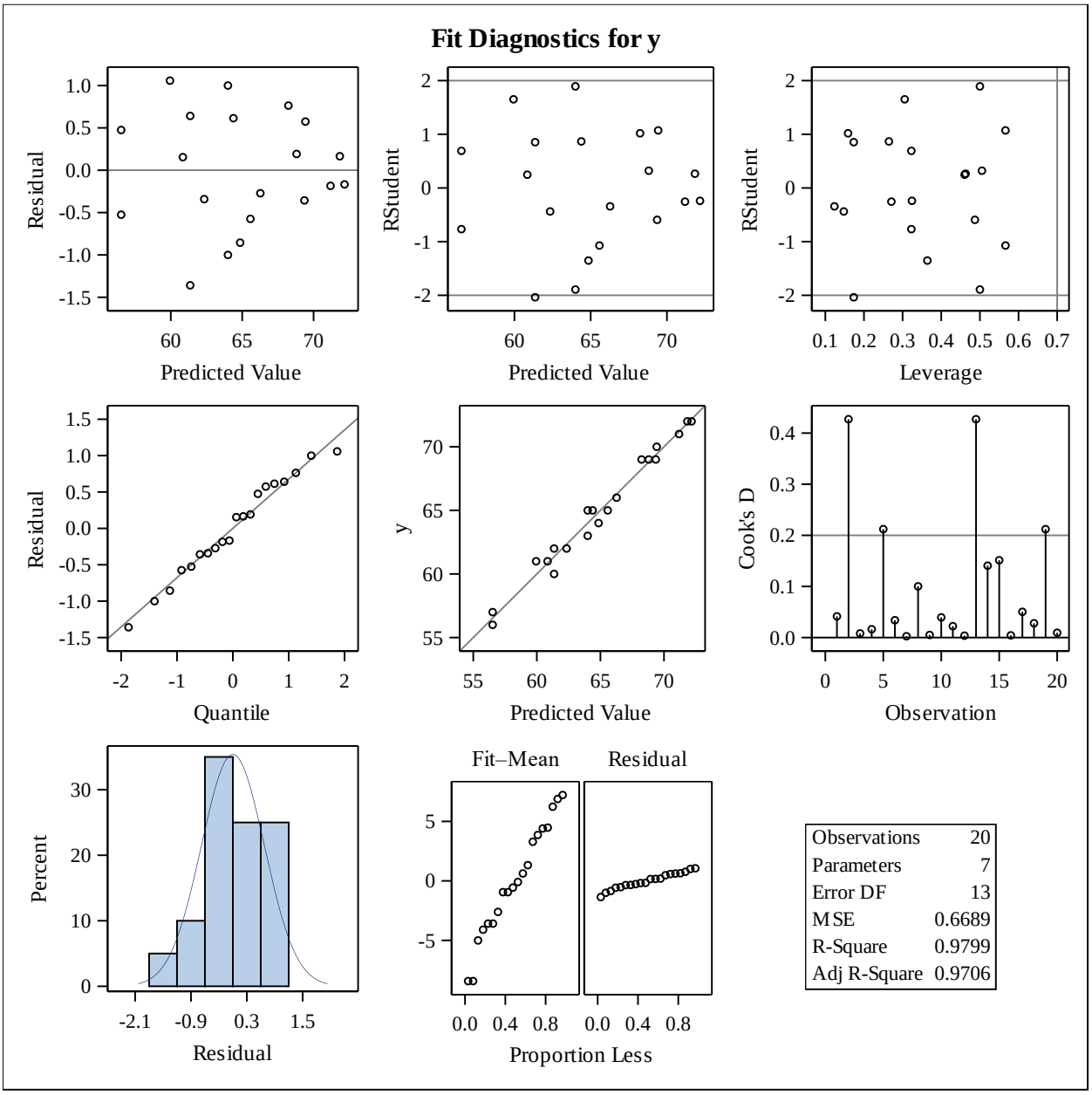
Root MSE	4.76994	**R-Square**	0.0541
Dependent Mean	64.95000	**Adj R-Sq**	0.0015
Coeff Var	7.34402		

Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	63.62500	1.68643	37.73	<.0001
x1	1	2.20833	2.17717	1.01	0.3239

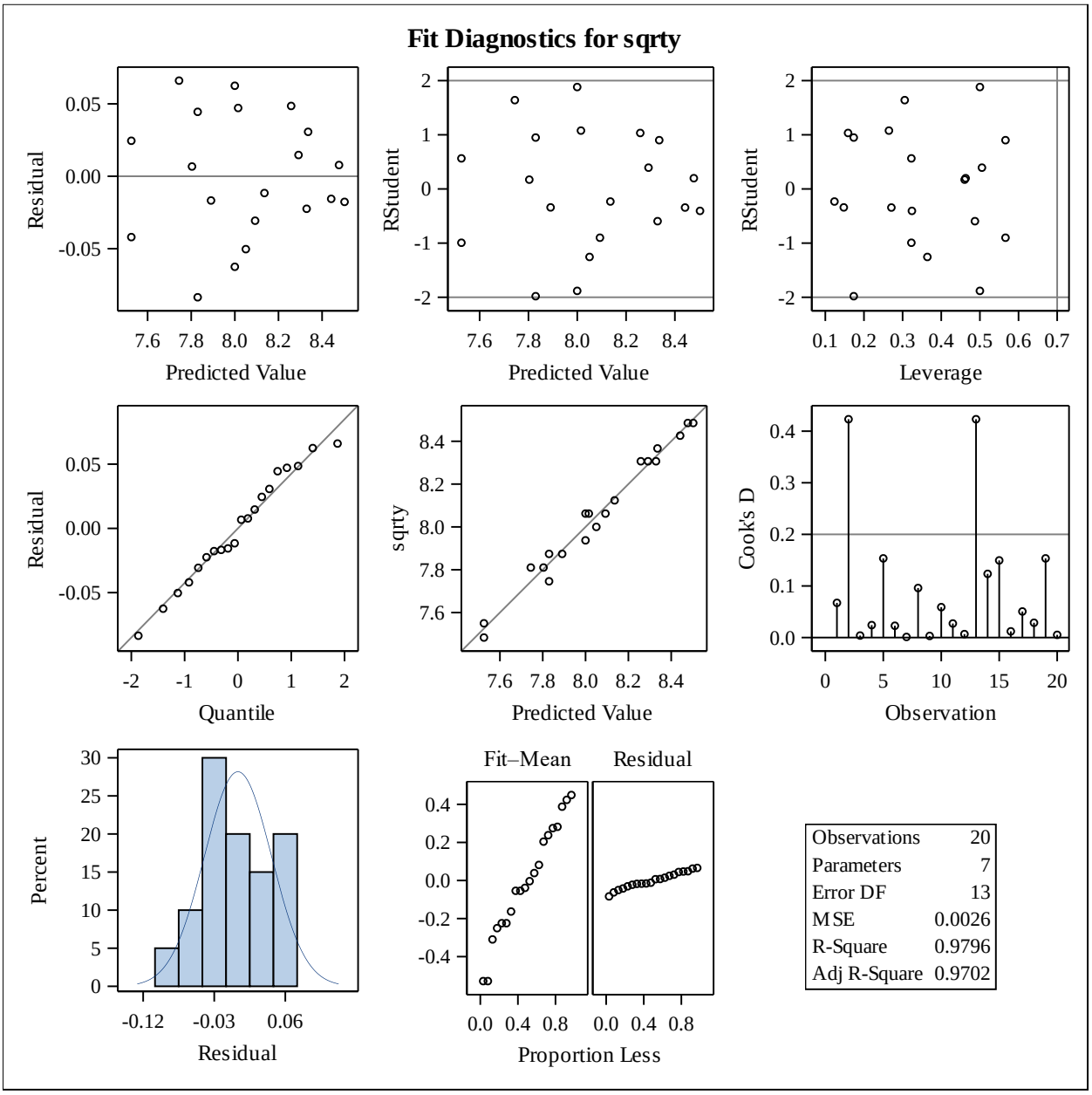
Partial SAS output for Model Selection Criteria for Models (1-4) in Part II

Number in Model	R-Square	C(p)	AIC	SBC	Variables in Model
Model (1)	0.9799	7.0000	-2.657	4.313	x1 x2 x3 x4 x5 x6
Model (2)	0.0541	596.2397	64.386	66.38	x1
Model (3)	0.9796	7.0000	-113.37	-106.4	x1 x2 x3 x4 x5 x6
Model (4)	0.0560	585.1446	-46.69	-44.70	x1

Diagnostic Plots for Models (1) in Part II



Diagnostic Plots for Models (3) in Part II



Part I

1. The treatments (WFH and CON) were randomly assigned to participants by the designer of this experiment.
2. The experimental units are employees (participants) and the experimental design of this study is a randomized complete block design (RCBD) or a matched pair design.
3. To test $\mu_{WFH} = \mu_{CON}$ versus $\mu_{WFH} \neq \mu_{CON}$,
   - **a.** The point estimate is 400 -360 = 40.
   - **b.** The standard deviation is about 23.

$$Var(d) = Var(X - Y) = Var(X) + Var(Y) - 2Cov(X, Y)$$
$$= 80^2 + 70^2 - 2 \times 0.12 \times 80 \times 70 = 9956$$

   And $Std(\bar{d}) = \sqrt{9956/19} \approx 23$

   - **c.** The *t*-test statistic = 40/23 = 1.74.
   - d. The degrees of freedom for the *t*-test is 19.
4. The assumptions for the paired sample *t*-test include: independence between teams, the differences between WFH and CON across teams follow an identical normal distribution.
5. The independence assumption is appropriate because the 20 teams were randomly sampled. To check whether the normality assumption is met, we need to take the differences between WFH and CON for each team and then check whether these 20 differences follow a normal distribution by using a normal probability plot or a test for normality such as the Shapiro-Wilk test.
6. Here is the completed ANOVA Table.

Source	**DF**
Team	19
Treatment	1
Error	19
Total	39

7. The linear model is

$$Y_{ij} = \mu + \beta_j + \alpha_i + \varepsilon_{ij}$$

   where μ corresponds to the overall mean, β_j corresponds to the block effect, α_i corresponds to the treatment effect, and the ε_{ij} are residual errors.
8. The terms μ and α_i are used to model the mean parameter of Y_{ij}. They are fixed effects as the treatment groups are specified and fixed for this study. The terms β_j and ε_{ij} are both random effects. Both teams (blocks) and employees were randomly selected for this study.
9. For the *F*-test:
   - **(a)** The *F*- statistic = $1.74^2 = 3.03$.
   - **(b)** The degrees of freedom are 1 for the numerator and 19 for the denominator.

10. The assumptions for the linear model include: μ and α_i are unknown fixed parameters, β_j are iid $N(0, \sigma_\beta^2)$ random variables, ε_{ij} are iid $N(0, \sigma^2)$ random variables, and the β_j are independent of the ε_{ij} for all i and j.

11. No, the assumptions for the *t*-test in **Problem 3** and the *F*-test in **Problem 9** are not the same. The assumptions for the *t*-test is more general because it does not need to specify the distribution of the block effect, and does not assume an additive model of the block and treatment effects.

12. The 1st, 2nd, and 20th rows of the design matrix **X** corresponding to employees 1, 2, and 20 (the observations listed in Table 2) for model (1).

$$\begin{vmatrix} 1 & 0 & 20 & 0 & 0 & 0 \\ 0 & 1 & 25 & 0 & 0 & 1 \\ & & & & & \\ 0 & 0 & 28 & 1 & 1 & 1 \end{vmatrix}$$

13. For employees with the same gender, age, marital status, having children or not, and having long commuting time ($\geq$ 90 minutes) or not, the mean satisfaction score for employees that work from home is 3.18 higher than the mean satisfaction score for employees that are in the control group.

14. Test to check whether work from home improved the employee's work satisfaction or not.

- **(a)** null: $\beta_1 \leq 0$ alternative: $\beta_1 > 0$
- **(b)** test statistic = 3.18/0.825 = 3.854,
- **(c)** df = 13,
- **(d)** *p*-value = 0.002/2=0.001
- **(e)** Yes, work from home improved the employee's work satisfaction score.

15. Based on the diagnostic plots for model (1), there are no obvious deviations from the normality and equal variance assumptions.

16. Test for $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$:

- **(a)** $F = \frac{(SSM_{model1} - SSM_{model2})/5}{MSE_{model1}} = \frac{(6*70.709 - 23.408)/5}{0.669} = 119.83$
- **(b)** The degrees of freedom are (5, 13)

17. The regression coefficient for the interaction term will NOT be estimable. The column in the design matrix for the interaction term will be exactly the same as the column for X_5.

18. Based on model selection criteria and model diagnostics on pages 8-10, either model (1) or model (3) would be okay. They both have smaller AIC and BIC than models (2) and (4) respectively, and the diagnostic plots show no obvious concerns for normality and equal variance. I would prefer using model (1) that is easier to interpret because it models the response on the original scale.

No tables of quantiles are provided with this exam. When necessary, express solutions in terms of quantiles defined as follows.

Let z_γ be the γ quantile of the standard normal distribution.

Let $t_{d,\gamma}$ be the γ quantile of the t distribution with d degrees of freedom.

Let $\chi^2_{d,\gamma}$ be the γ quantile of the chi-square distribution with d degrees of freedom.

Let $F_{d_1,d_2,\gamma}$ be the γ quantile of the F distribution with d_1 and d_2 degrees of freedom.

Part I

For $i = 1, \ldots, 50$, suppose $y_i = \mu + u_i + e_i$, where

$$u_1, \ldots, u_{50} \overset{iid}{\sim} N(0, \sigma^2) \text{ independent of } e_1, \ldots, e_{50} \overset{iid}{\sim} N(0, 1)$$

for some unknown parameters μ and $\sigma^2 > 0$. Suppose

$$y_1 = 18.8, \ \bar{y} = 17.2, \ \text{ and } \frac{1}{49}\sum_{i=1}^{50}(y_i - \bar{y})^2 = 3.5.$$

1. Provide a confidence interval for μ with coverage level 95%.

2. Provide a confidence interval for σ^2 with coverage level 95%.

3. Predict the value of u_1.

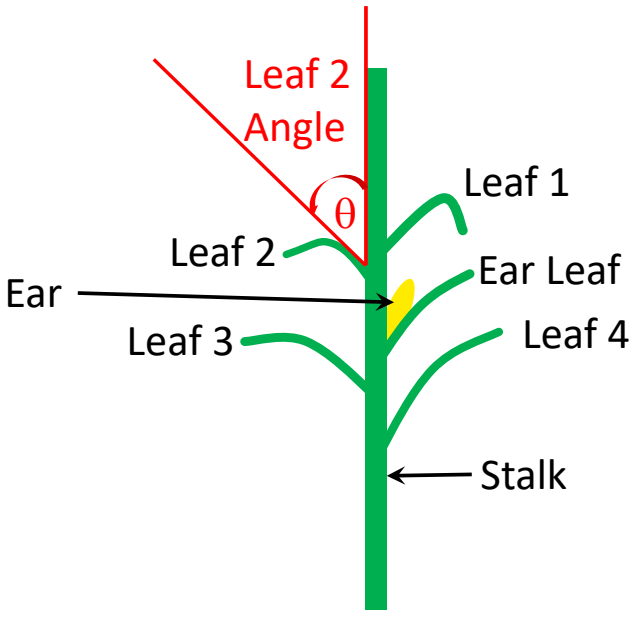
Part II

A maize plant has a vertical stalk and several leaves protruding from the stalk. The angle between each leaf and the stalk affects how much solar energy the plant can absorb. Researchers were interested in understanding how leaf angles vary across maize genotypes and how the spacing between adjacent plants might affect leaf angles.

Researchers conducted an experiment involving 8 maize genotypes planted in rows with 3 different plant spacings (10, 25, and 40 cm between adjacent plants in a row). A field was partitioned into 12 sections. Each section consisted of 16 rows. Within each section, the 8 maize genotypes were assigned, in a completely random way, to the 16 rows, with 2 rows for each genotype. The 3 plant spacings were assigned, in a completely random way, to the 12 sections, with 4 sections per spacing. Within each section, plants were spaced within rows according to the spacing randomly assigned to the section.

Eight weeks after planting, the angles of four leaves were measured for the plant at the center of each row. In particular, angles for the two leaves immediately above the ear and the two leaves immediately below the leaf closest to the ear were measured. These leaves are referred to from top down as leaves 1, 2, 3, and 4. Figure 1 depicts the angle measurement for leaf 2 of a prototypical plant. Leaf angles were measured in degrees with 0° representing an upright vertical leaf and 90° representing a horizontal leaf perpendicular to the stalk. The angle measurement for leaf 2 in Figure 1 is approximately 45°.

Figure 1: Angle Measurement for Leaf 2 of a Prototypical Maize Plant



4. The complete dataset includes 768 angle measurements (12 sections $\times$ 16 rows $\times$ 4 leaves per center plant in each row). In problem **4**, consider a dataset consisting of the 384 angle measurements for leaves 1 and 2 on each center plant. The design of the experiment giving rise to these 384 angle measurements can be characterized as a split-split-plot design. Consider the linear mixed-effects model you would fit to these data. Assume a model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ is appropriate, where $\boldsymbol{y}$ is the vector of 384 angle measurements, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are fixed matrices, $\boldsymbol{\beta}$ is a vector of parameters, and $[\boldsymbol{u}', \boldsymbol{e}']'$ is a multivariate normal random vector with mean zero and diagonal variance-covariance matrix.

a) Name the whole-plot factor.

b) Name the split-plot factor.

c) Name the split-split-plot factor.

d) State the numerator and denominator degrees of freedom for the F test for whole-plot factor main effects.

e) State the numerator and denominator degrees of freedom for the F test for split-plot factor main effects.

f) State the numerator and denominator degrees of freedom for the F test for split-split-plot factor main effects.

5. Now consider the complete dataset of 768 angle measurements. Suppose the angle measurements are arranged in a response vector $\boldsymbol{y}$ that is ordered first by section, then by row, and then by leaf number. Given the experiment's design and the structure of the dataset, consider the model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$ that you would fit to the data when completing **a)** and **b)** below.

a) Using Kronecker product notation, provide an expression for $\boldsymbol{Z}$.

b) Provide an expression for the variance-covariance matrix of the multivariate normal vector $[\boldsymbol{u}', \boldsymbol{e}']'$.

6. A few days after the angle measurements were taken, a straight-line wind storm, known as a derecho, blew through the field. Some of the maize plants were blown down while others remained standing. Researchers were interested in understanding which genotypes were most resistant to being blown down when plants were spaced 10 cm apart.

Let $i = 1, 2, 3, 4$ index the 4 sections with plants spaced 10 cm apart in each row. Let $j = 1, \ldots, 8$ index the 8 genotypes. Let $k = 1, 2$ index the two rows for any given genotype in any particular section. Let y_{ijk} be the number of plants left standing in the kth row for genotype j in section i after the derecho. A total of 21 plants were standing in each row prior to the derecho, so $y_{ijk} \in \{0, 1, \ldots, 21\}$ for all i, j, and k. Suppose

$$y_{ijk} \overset{ind}{\sim} \text{Binomial}(21, \pi_{ij}), \text{ where } \log\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \mu + \beta_i + \gamma_j, \tag{1}$$

for some unknown parameters μ, $\beta_1, \ldots, \beta_4$, and $\gamma_1, \ldots, \gamma_8$. Use the `R` code and output on pages 5 and 6 to complete **a)** through **g)** below.

a) What is the sum of the squared deviance residuals from the fit of model (1)?

b) Based on the fit of model (1), is there evidence of overdispersion? Explain your reasoning and support your answer with appropriate calculations.

c) Compute the test statistic you would use to test for differences among genotypes in resistance to being blown down.

d) State the null distribution of the test statistic in problem **c)**.

e) Use the fit of model (1) to estimate the probability that a plant of genotype 1 in section 1 remains standing.

f) Provide a confidence interval with coverage level approximately 95% for the probability that a plant of genotype 1 in section 1 remains standing.

g) Use the fit of model (1) to estimate, in any given section, the odds that a plant of genotype 8 remains standing divided by the odds that a plant of genotype 2 remains standing.

R Code and Output for Part II Problem 6

```
> section
 [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2
[27] 2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4
[53] 4 4 4 4 4 4 4 4 4 4 4 4
Levels: 1 2 3 4
> genotype
 [1] 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 1 1 2 2 3 3 4 4 5 5
[27] 6 6 7 7 8 8 1 1 2 2 3 3 4 4 5 5 6 6 7 7 8 8 1 1 2 2
[53] 3 3 4 4 5 5 6 6 7 7 8 8
Levels: 1 2 3 4 5 6 7 8
> y
 [1]  1  3  9  7  9 13 14 16  8  8 14  7  8  5  8 11  3
[18]  5 14  8  6 13 15 16 10 11 13 12 15 13  9 12  5  6
[35] 11  9 15 11 17 15  8 12 11 13 14  9 12 10  3  0 14
[52]  7 12 11 17 14 16  5  7 11  5 12  8 14
> summary(glm(cbind(y, 21 - y) ~ section + genotype,
+         family = binomial(link = logit)))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.69432  -0.67490   0.01723   0.66904   2.97684

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9942     0.2388  -8.350  < 2e-16 ***
section2      0.4473     0.1629   2.746  0.00603 **
section3      0.4869     0.1630   2.986  0.00282 **
section4      0.1982     0.1627   1.218  0.22317
genotype2     1.5905     0.2644   6.016 1.79e-09 ***
genotype3     1.8553     0.2645   7.015 2.31e-12 ***
genotype4     2.7564     0.2774   9.938  < 2e-16 ***
genotype5     1.5664     0.2644   5.923 3.16e-09 ***
genotype6     1.8071     0.2644   6.836 8.16e-12 ***
genotype7     1.6387     0.2643   6.200 5.63e-10 ***
genotype8     1.7108     0.2642   6.475 9.51e-11 ***

    Null deviance: 219.86  on 63  degrees of freedom
Residual deviance:  78.77  on 53  degrees of freedom
AIC: 312.17
```

R Code and Output for Part II Problem 6 (continued)

```
> summary(glm(cbind(y, 21 - y) ~ section,
+             family = binomial(link = logit)))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1200  -1.0010   0.0551   1.2926   3.2603

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.3242     0.1105  -2.933  0.00336 **
section2      0.4076     0.1554   2.623  0.00871 **
section3      0.4434     0.1555   2.852  0.00434 **
section4      0.1811     0.1555   1.165  0.24413

    Null deviance: 219.86  on 63  degrees of freedom
Residual deviance: 209.09  on 60  degrees of freedom
AIC: 428.5
```

1. $y_1, \ldots, y_{50} \overset{iid}{\sim} N(0, \eta^2)$, where $\eta^2 \equiv \sigma^2 + 1$. Thus, a confidence interval with coverage level 95% is

$$17.2 \pm t_{49,0.975}\sqrt{3.5/50}.$$

2.

$$\begin{aligned}
0.95 &= P\left[\chi^2_{49,0.025} \leq \frac{\sum_{i=1}^{50}(y_i - \bar{y})^2}{\eta^2} \leq \chi^2_{49,0.975}\right] \\
&= P\left[\chi^2_{49,0.025} \leq \frac{\sum_{i=1}^{50}(y_i - \bar{y})^2}{\sigma^2 + 1} \leq \chi^2_{49,0.975}\right] \\
&= P\left[\frac{1}{\chi^2_{49,0.975}} \leq \frac{\sigma^2 + 1}{\sum_{i=1}^{50}(y_i - \bar{y})^2} \leq \frac{1}{\chi^2_{49,0.025}}\right] \\
&= P\left[\frac{\sum_{i=1}^{50}(y_i - \bar{y})^2}{\chi^2_{49,0.975}} \leq \sigma^2 + 1 \leq \frac{\sum_{i=1}^{50}(y_i - \bar{y})^2}{\chi^2_{49,0.025}}\right] \\
&= P\left[\frac{\sum_{i=1}^{50}(y_i - \bar{y})^2}{\chi^2_{49,0.975}} - 1 \leq \sigma^2 \leq \frac{\sum_{i=1}^{50}(y_i - \bar{y})^2}{\chi^2_{49,0.025}} - 1\right] \\
&= P\left[49\frac{\frac{1}{49}\sum_{i=1}^{50}(y_i - \bar{y})^2}{\chi^2_{49,0.975}} - 1 \leq \sigma^2 \leq 49\frac{\frac{1}{49}\sum_{i=1}^{50}(y_i - \bar{y})^2}{\chi^2_{49,0.025}} - 1\right].
\end{aligned}$$

Thus, a confidence interval with coverage level 95% is

$$\left[49\frac{3.5}{\chi^2_{49,0.975}} - 1, 49\frac{3.5}{\chi^2_{49,0.025}} - 1\right] \Longleftrightarrow [1.44, 4.43].$$

3. To find the empirical best linear unbiased predictor, we find an expression for $E(u_1|y_1, \ldots, y_{50}) = E(u_1|y_1)$ and replace unknown parameters with their REML estimates:

$$\begin{aligned}
E(u_1|y_1) &= E(u_1) + \mathrm{Cov}(u_1, y_1)\mathrm{Var}^{-1}(y_1)(y_1 - E(y_1)) \\
&= 0 + \frac{\mathrm{Cov}(u_1, u_1 + e_1)}{\sigma^2 + 1}(y_1 - \mu) \\
&= 0 + \frac{\mathrm{Var}(u_1)}{\sigma^2 + 1}(y_1 - \mu) \\
&= 0 + \frac{\sigma^2}{\sigma^2 + 1}(y_1 - \mu) \\
&\approx \frac{2.5}{3.5}(18.8 - 17.2) \\
&\approx 1.14.
\end{aligned}$$

4. The natural split-split-plot linear mixed-effects model for the 384-observation dataset has fixed effects for all combinations of plant spacing, genotype, and leaf, as well as random effects for sections (whole-plot experimental units), rows (split-plot experimental units), and independent error terms corresponding to leaves. The random row effects induce correlation between the two observations (leaf 1 angle and leaf 2 angle) from a single plant because there is a one-to-one correspondence between rows and center plants for which angles were measured. The ANOVA for this model is

Source	DF
spacing	2
section(spacing)	9
geno	7
spacing $\times$ geno	14
geno $\times$ section(spacing) + row(geno, section, spacing)	63+96
leaf	1
spacing $\times$ leaf	2
geno $\times$ leaf	7
spacing $\times$ geno $\times$ leaf	14
leaf $\times$[section(spacing) + geno $\times$ section(spacing) + row(geno, section, spacing)]	9+63+96
corrected total	383

or, equivalently with less detail,

Source	DF
spacing	2
whole-plot error	9
geno	7
spacing $\times$ geno	14
split-plot error	159
leaf	1
spacing $\times$ leaf	2
geno $\times$ leaf	7
spacing $\times$ geno $\times$ leaf	14
split-split-plot error	168
corrected total	383

a) plant spacing

b) genotype

c) leaf

d) 2 and 9

e) 7 and 159

f) 1 and 168

5. **a**) Multiple solutions may receive full credit. Assuming random effects for sections and rows, $\boldsymbol{Z} = [\boldsymbol{I}_{12\times12} \otimes \mathbf{1}_{64\times1}, \boldsymbol{I}_{192\times192} \otimes \mathbf{1}_{4\times1}]$.

b) Multiple solutions could receive full credit. One solution is

$$\begin{bmatrix} \sigma_s^2\boldsymbol{I}_{12\times12} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_r^2\boldsymbol{I}_{192\times192} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{I}_{192\times192} \otimes \boldsymbol{\Sigma} \end{bmatrix},$$

where $\boldsymbol{\Sigma}$ is a 4×4 positive definite variance-covariance matrix. More specific choices for $\boldsymbol{\Sigma}$ could be proposed. $\boldsymbol{I}_{4\times4}$ is one possibility, but this would lead to equal correlations among all pairs of leaves from a single plant, which seems unlikely to be appropriate because of the ordering of leaves from top to bottom and because leaves 1 and 3 are on one side of the stalk while leaves 2 and 4 are on the other. An AR(1) structure for $\boldsymbol{\Sigma}$ might also be appropriate considering that the leaves are spatially ordered along the stalk. However, the leaves are not necessarily equally spaced, and as alluded to above, leaves on the same side of the stalk could be more correlated than leaves on different sides of the stalk. It is difficult to predict the most appropriate correlation structure. Thus, unstructured $\boldsymbol{\Sigma}$ may be the best place to start.

6. **a**) The sum of the squared deviance residuals is equal to the residual deviance 78.77.

b) Under the null hypothesis of no overdispersion, the residual deviance follows a chi-square distribution with degrees of freedom $n - p = 64 - 11 = 53$. The mean of a chi-square distribution with 53 degrees of freedom is 53, and the standard deviation of a chi-square distribution with 53 degrees of freedom is $\sqrt{2 \times 53} \approx 10.3$. Thus, the observed residual deviance is well over two standard deviations above the mean. More formally and with the aid of a computer, the probability that a chi-square random variable with degrees of freedom 53 exceeds 78.77 is 0.012. Thus, there is evidence of overdispersion.

c)

$$F = \frac{(209.09 - 78.77)/7}{78.77/53} \approx 12.53$$

d) The null distribution is approximated by an F distribution with 7 and 53 degrees of freedom.

e)

$$\frac{\exp(-1.9942)}{1 + \exp(-1.9942)} = \frac{1}{\exp(1.9942) + 1} \approx 0.12$$

f)

$$\left[\frac{1}{\exp(1.9942 + t_{53,0.975}\sqrt{(78.77/53)}0.2388) + 1}, \frac{1}{\exp(1.9942 - t_{53,0.975}\sqrt{(78.77/53)}0.2388) + 1}\right]$$

$$\approx [0.07, 0.20]$$

g)

$$\begin{aligned}
\left(\frac{\pi_{i8}}{1-\pi_{i8}}\right) \Big/ \left(\frac{\pi_{i2}}{1-\pi_{i2}}\right) &= \exp\left\{\log\left[\left(\frac{\pi_{i8}}{1-\pi_{i8}}\right) \Big/ \left(\frac{\pi_{i2}}{1-\pi_{i2}}\right)\right]\right\} \\
&= \exp\left\{\log\left(\frac{\pi_{i8}}{1-\pi_{i8}}\right) - \log\left(\frac{\pi_{i2}}{1-\pi_{i2}}\right)\right\} \\
&= \exp\left\{\gamma_8 - \gamma_2\right\}
\end{aligned}$$

Thus, the estimated odds ratio is $\exp(1.7108 - 1.5905) \approx 1.13$.

1 Background

The past year has seen many demonstrations by various groups of people supporting causes or protesting issues. Social scientists are sometimes interested in determining whether the occurrence of demonstrations on one topic (say, topic U) are associated with the occurrence of demonstrations on another topic (say, topic V). This question is concerned with the construction of statistical models to describe patterns in the occurrence of demonstrations on topic U (and on topic V), and whether there is any association among the occurrences of demonstrations on topic U and occurrences of demonstrations on topic V.

2 Data

Data are available from March 2020 to February 2021, with values reported monthly. We will restrict attention to cities with populations of $100,000$ or more, of which there were 314 in the U.S. as of July 2019 (according to Statista, a commercial provider of consumer data). Thus, we have available to us data at 314 geographic locations for 12 time windows each.
For each location and time window we have indicators of whether there were (indicator values 1) or were not (indicator values 0) demonstrations on topic U and demonstrations on topic V.

3 Basic Markov Random Field Models

We begin by considering the development of a model for only occurrences of demonstrations on topic U, and for only a **given time window**, based on binary conditional distributions in a Markov random field structure. Let $\{\boldsymbol{s}_i : i = 1, \ldots, n\}$ denote non-random random location variables (e.g., gps coordinates) such that $\boldsymbol{s}_i$ corresponds to a particular location within the United States. Corresponding to these locations, define a set of neighbors $\{N_i : i = 1, \ldots, n\}$ such that $\boldsymbol{s}_j \in N_i$ if $||\boldsymbol{s}_i - \boldsymbol{s}_j|| \leq d$ for some specified distance d.
Define random variables

$$Y(\boldsymbol{s}_i) = \begin{cases} 1 & \text{if a demonstration on topic U occurred at } \boldsymbol{s}_i, \\ 0 & \text{otherwise.} \end{cases}$$

A conditionally specified binary model results from assigning $Y(\boldsymbol{s}_i)$ the pmf, for $i = 1, \ldots, n$,

$$f(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : j \neq i\}) = \exp[A_i(\{y(\boldsymbol{s}_j) : j \neq i\}) - B_i(\{y(\boldsymbol{s}_j) : j \neq i\})], \tag{1}$$

where

$$A_i(\{y(\boldsymbol{s}_j) : j \neq i\}) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{j=1}^{n} \eta_{i,j}\{y(\boldsymbol{s}_j) - \kappa_j\}, \tag{2}$$

and

$$B_i(\{y(\boldsymbol{s}_j) : j \neq i\}) = \log(1 + \exp[A_i(\{y(\boldsymbol{s}_j) : j \neq i\})]).$$

We wish to fit the model (1) and (2) to data on topic U demonstrations.

1. What restrictions on parameters of the model in (1) and (2) would imply that the random variables $\{Y(\boldsymbol{s}_i) : i = 1, \ldots, n\}$ are mutually independent?

2. Using $[Y]$ as generic notation for the distribution of a random variable Y, recall that a Markov random field results from the assumption that, for $i = 1, \ldots, n$,

$$[Y(\boldsymbol{s}_i)|\{Y(\boldsymbol{s}_j) : j \neq i\}] = [Y(\boldsymbol{s}_i)|\{Y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in N_i\}].$$

What implications does this have for the set of parameters $\{\eta_{i,j} : i, j = 1, \ldots, n\}$ in (2)?

In order for the joint distribution of $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))^T$ to exist and be identified through use of what is called the negpotential function (the function $Q(\cdot)$ below), it is necessary that $\eta_{i,i} = 0$ for all i and $\eta_{i,j} = \eta_{j,i}$ in (2). If this is the case and we assume what is called pairwise-only dependence (technically necessary but not important to worry about for this question), then using Ω to denote the support of the joint distribution, the pmf of that distribution may be written as

$$g(\boldsymbol{y}) = \frac{\exp[Q(\boldsymbol{y})]}{\sum_{\boldsymbol{y}\in\Omega} \exp(\boldsymbol{y})}, \tag{3}$$

where

$$Q(\boldsymbol{y}) = \sum_{1\leq i\leq n} \left[\log\left(\frac{\kappa_i}{1-\kappa_i}\right) - \sum_{j\neq i} \eta_{i,j}\kappa_j\right] + \sum_{1\leq i<j\leq n}\sum \eta_{i,j} y(\boldsymbol{s}_i) y(\boldsymbol{s}_j). \tag{4}$$

Notational dependence of $g(\boldsymbol{y})$ and $Q(\boldsymbol{y})$ on parameters has been suppressed in (3). Those suppressed parameters appear on the right hand side of (4). In fact, the number of free parameters will have to be reduced to have a viable model. That will occur in the sequel, and is not our primary concern at this point.

3. For the purposes of estimation and inference, why is it not feasible, except in problems with very small sample size n (say $n \leq 10$ or so), to compute values of a log likelihood using (3) and (4) directly?

4. To accomplish estimation in a frequentist framework we might make use of composite likelihood theory for estimation. Write the explicit form (involving κs and ηs) of a composite log likelihood that could be used for estimation. Again, do not yet be concerned about the number of parameters appearing in your expressions.

5. An alternative form of composite likelihood would take the product of the joint conditional distributions of variables at locations and all of their neighbors, given all other variables. If we let $\boldsymbol{y}(N_i) = \{y(\boldsymbol{s}_j) : j \in N_i\}$ these conditional pmf's are

$$f(y(\boldsymbol{s}_i), \boldsymbol{y}(N_i) | \{y(\boldsymbol{s}_j) : j \neq i;\ \boldsymbol{s}_j \notin N_i\}). \tag{5}$$

Suppose that the largest neighborhood in the data set is $|N_i| = 4$. Demonstrate that the joint conditionals in (5) do not suffer the same problem as the complete joint distribution, which is the problem you identified in **Question 3**.

Hint: Consider that all marginal and conditional distributions one might want to define in this problem can be derived from the full joint pmf in (3).

We can now turn attention to the issue of reducing the number of free parameters in the model.

6. An extreme reduction in the number of free parameters is obtained by setting all of the κ_i equal to the same value κ and similarly setting all of the non-zero $\eta_{i,j}$ equal to the same value η. If this is done, what does the parameter η represent **in terms of the occurrence probabilities of demonstrations at our locations**?

Hint: Recall our previous definition of neighborhoods, the relation between the probabilities of concern and the functions A_i in (2), and what form those functions take under the independence model of **Question 1**.

7. Let $\boldsymbol{x}_i = \{x_{i,j} : i = 1, \ldots, n;\ j = 1, \ldots, p\}$ represent a set of p covariates measured at each of our n locations. These covariates represent characteristics of locations that might influence the occurrence probability of a demonstration on topic U, such as demographic

characteristics of a location's population. We will treat these covariates as non-random quantities in our models. How would you incorporate information attached to these covariates into the model of (1) and (2)? Be explicit, giving a formula.

Hint: Such covariates might well be important regardless of whether or not occurrences of demonstrations are independent among locations.

4 Models over Time

Now consider taking a basic Markov random field model for a static point in time and extending it to incorporate all 12 time windows for which data are available. Two possibilities immediately suggest themselves for accomplishing this goal.

The first approach to incorporating temporal structure is to simply extend the Markov random field structure of a basic model by specifying that each city at one time point is included in the neighborhood of that same city at the next point in time. Notationally, this can be accomplished by changing the indexing so that the pair $(\boldsymbol{s}_i, t)$ indicates a city $\boldsymbol{s}_i$ at time window t. We take responses from the first and the last time windows to be given values so they will not contribute response random variables to the model but will be considered as conditioning values only. Let the values of the first time window be denoted as $\{y(\boldsymbol{s}_i, 0) : i = 1, \ldots, n\}$ and the values of the last time window as $\{y(\boldsymbol{s}_i, T+1) : i = 1, \ldots, n\}$. Our response random variables are then $\{Y(\boldsymbol{s}_i, t) : i = 1, \ldots, n;\ t = 1, \ldots, T\}$ with $T = 10$ for our data. We then extend the definition of neighborhoods as, for $i = 1, \ldots, n$ and $t = 1, \ldots, T$,

$$N_{i,t} = \{(\boldsymbol{s}_j, m) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| \leq d;\ m = t \pm 1\}. \tag{6}$$

The model is completed by replacing (2) with

$$\begin{aligned} A_{i,t}(\{y(\boldsymbol{s}_j, t), y(\boldsymbol{s}_i, m) : j \neq i;\ m = t \pm 1\}) &= \log\left(\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right) + \sum_{j=1}^{n} \eta_{i,j}\{y(\boldsymbol{s}_j, t) - \kappa_{j,t}\} \\ &+ \sum_{m \in \{t \pm 1\}} \eta_{i,m}\{y(\boldsymbol{s}_i, m) - \kappa_{i,m}\}, \end{aligned} \tag{7}$$

where $\eta_{i,i} = 0$ for all i, $\eta_{i,j} = \eta_{j,i}$, and

$$B_{i,t}(\{y(\boldsymbol{s}_j, t), y(\boldsymbol{s}_i, m) : j \neq i;\ m = \pm 1\}) = \log(1 + \exp[A_i(\{y(\boldsymbol{s}_j, m) : j \neq i;\ m = \pm 1\})]).$$

Another possible approach for adding temporal structure to the model is to incorporate an autoregressive-like term into the large-scale model component. In this case, (2) becomes

$$A_{i,t}(\{y(\boldsymbol{s}_j,t) : j \neq i\}) = \log\left(\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right) + \sum_{j=1}^{n} \eta_{i,j}\{y(\boldsymbol{s}_j) - \kappa_{j,t}\}, \tag{8}$$

where now

$$\log\left(\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right) = \log\left(\frac{\lambda}{1-\lambda}\right) + \gamma\left[\log\left(\frac{\kappa_{i,t-1}}{1-\kappa_{i,t-1}}\right) - \log\left(\frac{\lambda}{1-\lambda}\right)\right] + \epsilon_t, \tag{9}$$

with the ϵ_t iid following a normal distribution with mean 0 and variance σ^2, for $t = 1, \ldots, T$ and $i = 1, \ldots, n$.

8. Comment on the dependence among **response** variables in the model of (7) and that of (8) and (9). Do both of these models allow for dependence between $Y(\boldsymbol{s}_i, t)$ and $Y(\boldsymbol{s}_i, t-1)$? Do both allow for dependence between $Y(\boldsymbol{s}_i, t)$ and $Y(\boldsymbol{s}_i, t+1)$?

9. For the model given by (8) and (9), first set all $\eta_{i,j} = 0$ and set all $\kappa_{i,t} = \kappa$. Assume that

$$E\left[\log\left(\frac{\kappa_0}{1-\kappa_0}\right)\right] = \log\left(\frac{\lambda}{1-\lambda}\right).$$

What can you say about the expected values of $Y(\boldsymbol{s}_i, t)$ under these restrictions?

Hint: If we set all $\eta_{i,j} = 0$ and all $\kappa_i = \kappa$ in (2), the expected value of $Y(\boldsymbol{s}_i)$ is κ.

5 Models for Two Types of Responses

We now add responses associated with demonstrations on topic V. For this purpose, define the random variables

$$Z(\boldsymbol{s}_i, t) = \begin{cases} 1 & \text{if a demonstration on topic V occurred at } \boldsymbol{s}_i \text{ in time window } t, \\ 0 & \text{otherwise.} \end{cases}$$

Development of a model for this $\boldsymbol{Z}$ process would be analogous to that of developing a model for the $\boldsymbol{Y}$ process that has been the focus of this question so far.

We now have random variables associated with both demonstrations on topic U as $Y(\boldsymbol{s}_i, t)$ and with demonstrations on topic V as $Z(\boldsymbol{s}_i, t)$, both for $i = 1, \ldots, n$ and $t = 1, \ldots, T$. As stated in the Background section of this question, we would like to develop a model that allows a relation

between $Y(\boldsymbol{s}_i, t)$ and $Z(\boldsymbol{s}_i, s)$ for one or more values of s. For simplicity, take $s \in \{t-1, t, t+1\}$ and consider the task first to be formulating a model that relates one or more of $Z(\boldsymbol{s}_i, s)$ to $Y(\boldsymbol{s}_i, t)$.

We will further reduce the complexity of this problem by temporarily dropping terms in the model that relate components of the $\boldsymbol{Y}$ process in space or over time, and similarly for the $\boldsymbol{Z}$ process.

10. Our problem is now similar to that of relating $Y(\boldsymbol{s}_i, t)$ to $Y(\boldsymbol{s}_i, t+1)$ and/or $Y(\boldsymbol{s}_i, t-1)$ in the previous section. That is, there are again at least two overall structures that suggest themselves for use in this task, an extended Markov random field structure and an autoregressive (AR) structure in the large-scale model component. Without including any terms to relate the $\boldsymbol{Y}$s spatially or over time, or the $\boldsymbol{Z}$s spatially or over time, write the models that would result from these two approaches. Introduce new notation for previously unused parameters if necessary.

Give one characteristic of the Markov random field approach that might make it more attractive than the alternative AR approach for modeling associations between occurrences of demonstrations on topic U ($\boldsymbol{Y}$s) and occurrences of demonstrations on topic V (the $\boldsymbol{Z}$s). Give one characteristic of the Markov random field approach that might make it less attractive than the alternative AR approach.

These are a sketch of the answers hoped for. For some of the questions, other possibilities might exist that would be entirely adequate if they are both technically correct and logically consistent.

1. What is called an independence model results from taking $\eta_{i,j} = 0$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

2. The implication is that, for $i = 1, \ldots, n$, $\eta_{i,j} = 0$ unless $\boldsymbol{s}_j \in N_i$.

3. The difficulty is that with even moderate n, the size of the joint support Ω (which is 2^n) becomes too large for the denominator of expression (3) in the question to be computed efficiently, especially in an iterative estimation algorithm in which the likelihood must be computed repeatedly.

4. Let the events of a composite likelihood be defined as $Y(\boldsymbol{s}_i)$ given all other values. Then Besag's original pseudo-likelihood is a composite log likelihood,

$$\begin{aligned}\ell_c(\boldsymbol{\kappa}, \boldsymbol{\eta}) &= \sum_{i=1}^{n} \log[f(y(\boldsymbol{s}_i)|\{y(\boldsymbol{s}_j) : \boldsymbol{s}_j \in N_i\})] \\ &= \sum_{i=1}^{n} \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{\boldsymbol{s}_j \in N_i} \eta_{i,j}\{y(\boldsymbol{s}_j) - \kappa_j\}.\end{aligned}$$

5. Consider a particular set $\mathcal{A} = \{y(\boldsymbol{s}_i), \boldsymbol{y}(N_i)\}$ for which $|N_i| = 4$, and let $\mathcal{A}^c = \{y(\boldsymbol{s}_j) : j \neq i;\ j \notin N_i\}$ denote the set of all values not in $\mathcal{A}$. Note that $|\mathcal{A}| = 5$ and $|\mathcal{A}^c| = n - 5$.

Using $g(\cdot)$ to denote the joint pmf of $\boldsymbol{y}$ (as in the question), the conditional pmf we need is

$$f(\mathcal{A}|\mathcal{A}^c) = \frac{g(\mathcal{A}, \mathcal{A}^c)}{g(\mathcal{A}^c)}. \tag{1}$$

Here,

$$g(\mathcal{A}, \mathcal{A}^c) = \frac{\exp[Q(\mathcal{A}, \mathcal{A}^c)]}{\sum_{\boldsymbol{y} \in \mathcal{A}} \sum_{\boldsymbol{t} \in \mathcal{A}^c} \exp[Q(\boldsymbol{y}, \boldsymbol{t})]}, \tag{2}$$

and

$$g(\mathcal{A}^c) = \frac{\sum_{\boldsymbol{y} \in \mathcal{A}} \exp[Q(\mathcal{A}, \mathcal{A}^c)]}{\sum_{\boldsymbol{y} \in \mathcal{A}} \sum_{\boldsymbol{t} \in \mathcal{A}^c} \exp[Q(\boldsymbol{y}, \boldsymbol{t})]}. \tag{3}$$

Substitution of (2) and (3) into (1) gives

$$f(\mathcal{A}|\mathcal{A}^c) = \frac{\exp[Q(\mathcal{A}, \mathcal{A}^c)]}{\sum_{\boldsymbol{y} \in \mathcal{A}} \exp[Q(\mathcal{A}, \mathcal{A}^c)]}. \tag{4}$$

The summation in the denominator of (4) contains only $2^5 = 32$ terms, which can be easily computed, even as part of an iterative optimization algorithm.

6. Following the hint, we have that the natural parameter function for our Markov random field model may now be written as

$$A_i(\{y(\boldsymbol{s}_j) : j \neq i\}) = \log\left(\frac{\kappa}{1-\kappa}\right) + \sum_{\boldsymbol{s}_j \in N_i} \eta\{y(\boldsymbol{s}_j) - \kappa\}, \tag{5}$$

and for the independence model as

$$A_i = \log\left(\frac{\kappa}{1-\kappa}\right). \tag{6}$$

Assuming that $\eta > 0$, this parameter thus represents the change in the natural parameter functions A_i from the independence model caused by a neighboring value of 1 (increase in A_i) or 0 (decrease in A_i). Since the probability that a location has a realized value of 1 is monotone in A_i (formally is $\exp(A_i)/[1 + \exp(A_i)]$), η is related to the increase or decrease in the probability a demonstration on topic U occurs at location $\boldsymbol{s}_i$ caused by the occurrence or lack thereof at neighboring locations.

7. Particularly given the hint, we would like to incorporate covariate information into the large-scale model structure. Borrowing the notation of a link function from basic generalized linear models, a natural way to accomplish this is, with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ being an unknown regression parameter,

$$\log\left(\frac{\kappa_i}{1-\kappa_i}\right) = \boldsymbol{x}_i^T \boldsymbol{\beta}.$$

8. Both models result in serial dependence among values of the response variables, and both imply dependence among $Y(\boldsymbol{s}_i, t)$ and both $Y(\boldsymbol{s}_i, t-1)$ and $Y(\boldsymbol{s}_i, t+1)$. This is true in the joint distribution (which is primarily what is wanted in the answer to this question), and is also true in the full conditional distributions. In particular, the full conditional distribution of $Y(\boldsymbol{s}_i, t)$ will depend on both $Y(\boldsymbol{s}_i, t-1)$ and $Y(\boldsymbol{s}_i, t+1)$ even in the model of expressions (8) and (9), because that model explicitly gives only directional conditionals. It can be shown, although that was not expected in answers to this question, that the traditional Markov property in time implies full conditional distributions depend on variables both one step before and one step after a given point in time.

9. Setting all $\eta_{i,j} = 0$ and $\kappa_{i,t} = \kappa_t$ in expression (8) of the question gives

$$A_i = \log\left(\frac{\kappa_t}{1-\kappa_t}\right),$$

where

$$\log\left(\frac{\kappa_t}{1-\kappa_t}\right) = \log\left(\frac{\lambda}{1-\lambda}\right) + \gamma\left[\log\left(\frac{\kappa_{t-1}}{1-\kappa_{t-1}}\right) - \log\left(\frac{\lambda}{1-\lambda}\right)\right] + \epsilon_t.$$

Following the hint, the conditional expected value of $Y(\boldsymbol{s}_i, t)$ given κ_t is then

$$E[Y(\boldsymbol{s}_i, t)|\kappa_t] = \kappa_t,$$

so the marginal expectation is $E[Y(\boldsymbol{s}_i, t)] = E(\kappa_t)$. Now, the logit of the κ_t follows an AR(1) process and the expected value of the logit of κ_0 is the logit of λ. Then by recursion it is easy to show that, for $t = 1, \ldots, T$,

$$E\left[\log\left(\frac{\kappa_t}{1-\kappa_t}\right)\right] = \log\left(\frac{\lambda}{1-\lambda}\right).$$

The function $\log[x/(1-x)]$ is concave for $x \leq 0.5$ and convex for $x \geq 0.5$. Then Jensen's inequality implies that

$$\begin{aligned} E(\kappa_t) &\geq \lambda \text{ for } \kappa_t \leq 0.50 \\ E(\kappa_t) &\leq \lambda \text{ for } \kappa_t \geq 0.50 \end{aligned}$$

with the inequalities being strict except at $\kappa_t = 0.50$. These same inequalities will hold for the $Y(\boldsymbol{s}_i, t)$ as

$$\begin{aligned} E[Y(\boldsymbol{s}_i, t)] &\geq \lambda \text{ for } \kappa_t \leq 0.50 \\ E[Y(\boldsymbol{s}_i, t)] &\leq \lambda \text{ for } \kappa_t \geq 0.50 \end{aligned}$$

10. Take the natural parameter functions of the $Y(\boldsymbol{s}_i, t)$ to be $A_{y,i}$, and those for the $Z(\boldsymbol{s}_i, t)$ to be $A_{z,i}$. Let $\mathcal{S} = \{t-1, t, t+1\}$. Then, without any spatial or temporal structure within the $\boldsymbol{Y}$ process, the conditional pmf's of the $Y(\boldsymbol{s}_i, t)$ are

$$f(y(\boldsymbol{s}_i, t)|\{z(\boldsymbol{s}_i, s) : s \in \mathcal{S}) = \exp[A_{y,i}(\{z(\boldsymbol{s}_i, s) : s \in \mathcal{S}\}) - B_i(\{z(\boldsymbol{s}_i, s) : s \in \mathcal{S}\})],$$

and, without any spatial or temporal structure in the $\boldsymbol{Z}$ process, those of the $Z(\boldsymbol{s}_i, t)$ are

$$g(z(\boldsymbol{s}_i, t)|\{y(\boldsymbol{s}_i, s) : s \in \mathcal{S}) = \exp[A_{z,i}(\{y(\boldsymbol{s}_i, s) : s \in \mathcal{S}\}) - B_i(\{y(\boldsymbol{s}_i, s) : s \in \mathcal{S}\})].$$

To extend a Markov random field structure to this situation, define the neighborhood of $(\boldsymbol{s}_i, t)$ as

$$N_{i,t} = \{(\boldsymbol{s}_i, s) : s \in \mathcal{S}\}.$$

Let $\boldsymbol{z}(N_{i,t}) = \{z(\boldsymbol{s}_i, s) : \ s \in \mathcal{S}\}$, and $\boldsymbol{y}(N_{i,t}) = \{y(\boldsymbol{s}_i, s) : \ s \in \mathcal{S}\}$. Then model the natural parameter functions as

$$\begin{aligned} A_{y,i}(\boldsymbol{z}(N_{i,t})) = \log\left(\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right) + \eta_1\{z(\boldsymbol{s}_i, t-1) - \psi_{i,t-1}\} + \\ \eta_2\{z(\boldsymbol{s}_i, t) - \psi_{i,t}\} + \eta_3\{z(\boldsymbol{s}_i, t+1) - \psi_{i,t+1}\}, \\ A_{z,i}(\boldsymbol{y}(N_{i,t})) = \log\left(\frac{\psi_{i,t}}{1-\psi_{i,t}}\right) + \eta_3\{y(\boldsymbol{s}_i, t-1) - \kappa_{i,t-1}\} + \\ \eta_2\{y(\boldsymbol{s}_i, t) - \kappa_{i,t}\} + \eta_1\{y(\boldsymbol{s}_i, t+1) - \kappa_{i,t+1}\}. \end{aligned}$$

Note, the placements of η_1 and η_3 (and the fact there can be only two such parameters for cross-dependencies) in the above expressions are necessary in this modeling approach, but it is not expected that students taking the exam will recognize that. That is, that technical detail is not the important part of the answer. What is important is extending the neighborhood structure and then modeling dependencies as additive components in the natural parameter function.

To formulate a model with an AR structure in the large-scale model component we might take

$$\begin{aligned} A_{y,i}(\kappa_{i,t}) &= \log\left(\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right), \\ A_{z,i}(\psi_{i,t}) &= \log\left(\frac{\psi_{i,t}}{1-\psi_{i,t}}\right), \end{aligned}$$

where

$$\begin{aligned} \log\left(\frac{\kappa_{i,t}}{1-\kappa_{i,t}}\right) &= \log\left(\frac{\lambda}{1-\lambda}\right) + \gamma_\kappa\left[\log\left(\frac{\psi_{i,t-1}}{1-\psi_{i,t-1}}\right) - \log\left(\frac{\xi}{1-\xi}\right)\right] + \epsilon_t, \\ \log\left(\frac{\psi_{i,t}}{1-\psi_{i,t}}\right) &= \log\left(\frac{\xi}{1-\xi}\right) + \gamma_\psi\left[\log\left(\frac{\kappa_{i,t-1}}{1-\kappa_{i,t-1}}\right) - \log\left(\frac{\lambda}{1-\lambda}\right)\right] + \epsilon_t, \end{aligned}$$

and the $\epsilon_{i,t}$ are iid following a common normal distribution with expected value 0 and variance σ^2.

One aspect of the Markov random field approach that makes it perhaps more attractive than the alternative AR approach is that cross-dependencies between, for example $Y(\boldsymbol{s}_i, t)$ and $Z(\boldsymbol{s}_i, t-1)$, are modeled more explicitly than in the AR approach, in which those associations occur in the joint distribution only as implied by the modeled dependence between $\kappa(\boldsymbol{s}_i, t)$ and $\psi(\boldsymbol{s}_i, t)$. One aspect of the Markov random field approach that might make it less attractive than the alternative AR approach is that the symmetry requirement for dependence among neighbors takes the effect of a difference in $z(\boldsymbol{s}_i, t) - \psi_{i,t}$ on $y(\boldsymbol{s}_i, t)$ to be the same as a difference in $y(\boldsymbol{s}_i, t) - \kappa_{i,t}$ on $z(\boldsymbol{s}_i, t)$. That is, it takes the occurrences of demonstrations on topic U to have the same effect on the occurrences of demonstrations on topic V as the other way around. The AR structure allows these to be different because γ_κ may not equal γ_ψ.