A random sample of 56 customers from a large clothing retailer was selected. For each customer, the values of the following variables were obtained from store records.

Customer ID: Customer identification number for record keeping purposes.
Purchase: The net dollar amount spent by customer in his or her last purchase from this retailer, rounded to the nearest dollar.
Time: Number of months since last purchase.
Number12: Number of purchases in the last 12 months.
Total12: Net total dollar amount of purchases in the last 12 months, rounded to the nearest dollar.
Number24: Number of purchases in the last 24 months.
Total24: Net total dollar amount of purchases in the last 24 months, rounded to the nearest dollar.
Card: Whether or not the customer has a private label store credit card (1 = Yes, 0 = No).

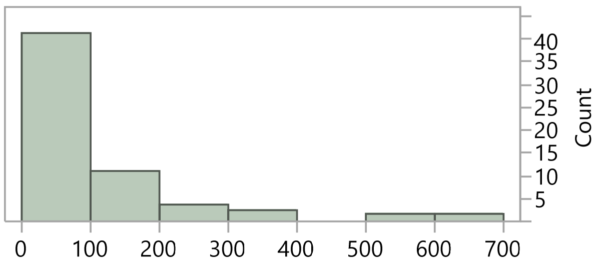
The table on the last page of this problem contains the values of the above variables for all 56 customers in the random sample.

The store manager is interested in predicting the net dollar amount a customer will spend on his or her next purchase using the explanatory variables: Time, Number12, Total12, Number24, Total24, and Card.

Part I

1. Below are graphical and numerical summaries of the response variable, Purchase. Describe the distribution of the variable. Based on your description, indicate any potential issues with using this response variable in a multiple linear regression analysis.

Purchase



Quantiles

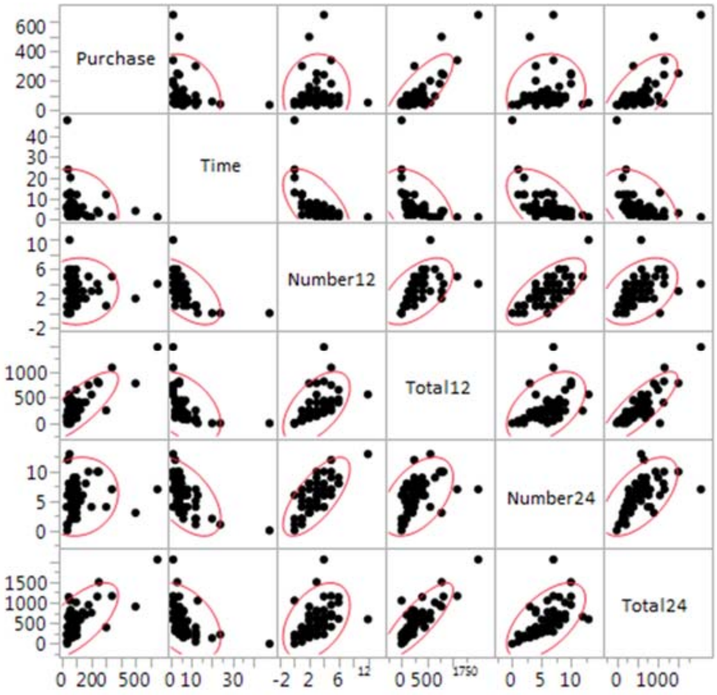
100.0%	maximum	650
75.0%	quartile	100
50.0%	median	71
25.0%	quartile	50
0.0%	minimum	30

Summary Statistics

Mean	108.28571
Std Dev	112.18843
N	56

2. Below is a correlation matrix and a scatterplot matrix of the quantitative variables in this analysis. Summarize the relationship between the response variable Purchase and the quantitative explanatory variables in the analysis. Which explanatory variables have the strongest and weakest relationship with the response variable, Purchase?

	Purchase	**Time**	**Number12**	**Total12**	**Number24**	**Total24**
Purchase	1.0000					
Time	-0.2208	1.0000				
Number12	0.0516	-0.5838	1.0000			
Total12	0.8037	-0.4539	0.5559	1.0000		
Number24	0.1017	-0.5491	0.7100	0.4849	1.0000	
Total24	0.6773	-0.4324	0.4215	0.8275	0.5962	1.0000



3. The output from Problem **2** also gives you a summary of the pairwise relationships among the quantitative explanatory variables. Summarize these relationships and indicate any potential issues these relationships might raise in the multiple linear regression analysis.
4. Below are numerical summaries of the response variable, Purchase, between the two groups given by the explanatory variable Card. Summarize the differences in the distributions of the response variable, Purchase, between the two groups in the random sample. Is there a significant difference in the mean purchase amounts between the population of customers who have a card and those who don't? Write "Yes" or "No" and explain your answer.

Card	**Number**	**Mean**	**Std. Dev.**	**Min**	**25%**	**Median**	**75%**	**Max**
0 (no)	37	75.973	47.324	30	50	65	85	300
1 (yes)	19	171.211	166.292	39	58	100	240	650

Part II

For this part of the question, we will consider a multiple linear regression model to predict the variable Purchase from all six explanatory variables.

5. Below is a table with the Type I Sums of Squares for the six explanatory variables in the model. Conduct a hypothesis test for the significance of the overall model in predicting the response variable, Purchase.

Variable	Type I SS
Time	33753
Number12	6278
Total12	565766
Number24	2987
Total24	74
Card	1679

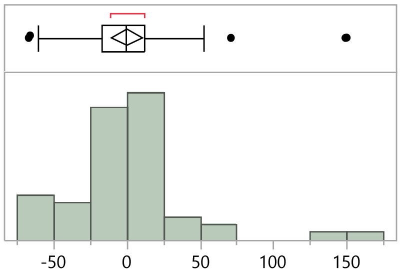
6. Below is the table of parameter estimates for this multiple linear regression model. Use the information in the table to conduct hypothesis tests for the significance of each of Total12 and Total24 in the model. Explain why the outcomes of these tests make sense given the context of the data.

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	111.56414	21.92715	5.09	<.0001*	67.499876	155.6284	.
Time	-1.345963	0.971053	-1.39	0.1720	-3.297368	0.6054422	1.6562382
Number12	-32.35354	5.18787	-6.24	<.0001*	-42.77895	-21.92812	3.0825433
Total12	0.4296834	0.041325	10.40	<.0001*	0.3466377	0.5127291	4.5409476
Number24	-5.173593	3.619661	-1.43	0.1593	-12.44757	2.1003879	3.2568561
Total24	0.0017562	0.03185	0.06	0.9562	-0.062248	0.0657608	4.8081163
Card[1]	14.624409	14.57577	1.00	0.3206	-7.3333487	21.95776	1.5994424

7. The table in Problem **6** also includes information about the Variance Inflation Factor (VIF) for each explanatory variable. Explain why the values for VIF are the highest for the Total12 and Total24 explanatory variables.

8. Below are graphical and numerical summaries for the distribution of residuals from the multiple linear regression model and residual plots for each explanatory variable. Use this information to indicate any model assumptions that might be violated.

Distribution of Residuals



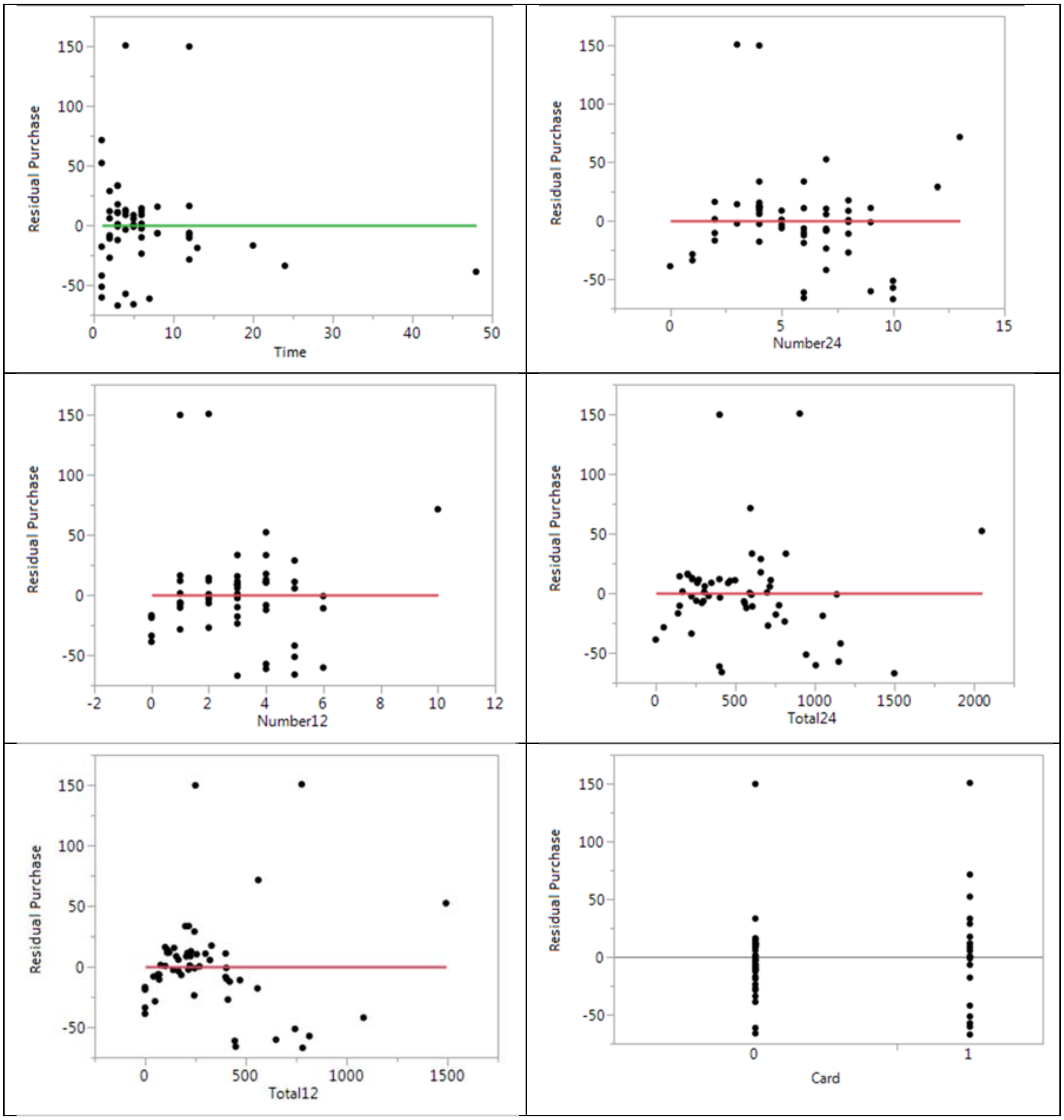
Quantiles

100.0%	maximum	150.62932217
75.0%	quartile	11.662213361
50.0%	median	-1.060655186
25.0%	quartile	-17.64620711
0.0%	minimum	-67.16053795

Summary Statistics

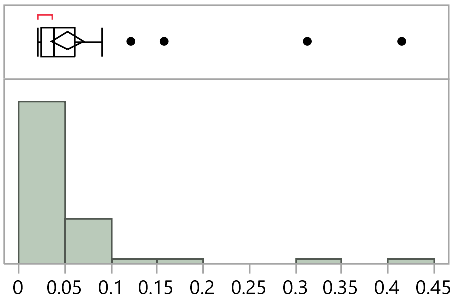
Std Dev	40.162288

Residual Plots



9. The distributions of hat values (h_{ii}) and Cook’s D values for this multiple linear regression model are summarized below. Which of these values indicate any observations with either high leverage, high influence, or both? Make sure to describe the criteria used.

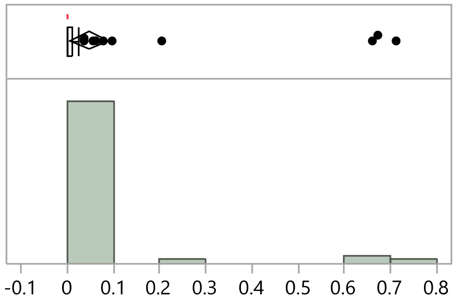
Hat Value



Quantiles

100.0%	maximum	0.4164279792
75.0%	quartile	0.0607283989
50.0%	median	0.0367798085
25.0%	quartile	0.0240833649
0.0%	minimum	0.0192710794

Cook's D Value



Quantiles

100.0%	maximum	0.714817036
75.0%	quartile	0.0100064384
50.0%	median	0.0011421079
25.0%	quartile	0.000330323
0.0%	minimum	4.560194e-7

Part III

For this part of the question, we will consider two multiple linear regression models to predict the response variable Purchase.

Model **1** with Number12, Total12, and Card
Model **2** with Number12 and Total12

10. The Sum of Squares for Error for Model **1** is 88092.8. Use this number and information given in Problem **2** for the Total Sums of Squares to complete the ANOVA Table and conduct a test of significance of the model in predicting the response variable Purchase.

11. Below is the parameter estimates table for Model **1**. Give an interpretation of the estimated coefficient for Card in the model. Is this coefficient statistically significant in this model? Explain your answer.

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	77.674349	12.23607	6.35	<.0001*
Number12	-34.5807	3.591921	-9.63	<.0001*
Total12	0.4366897	0.025814	16.92	<.0001*
Card[1]	8.2016756	13.52815	0.61	0.5470

12. In Problem **4**, you were asked to compare the response variable, Purchase, between the two groups defined by values of the explanatory variable Card. How does your Problem **4** analysis compare to your answer to Problem **11**? Explain any similarities or differences in your conclusions.

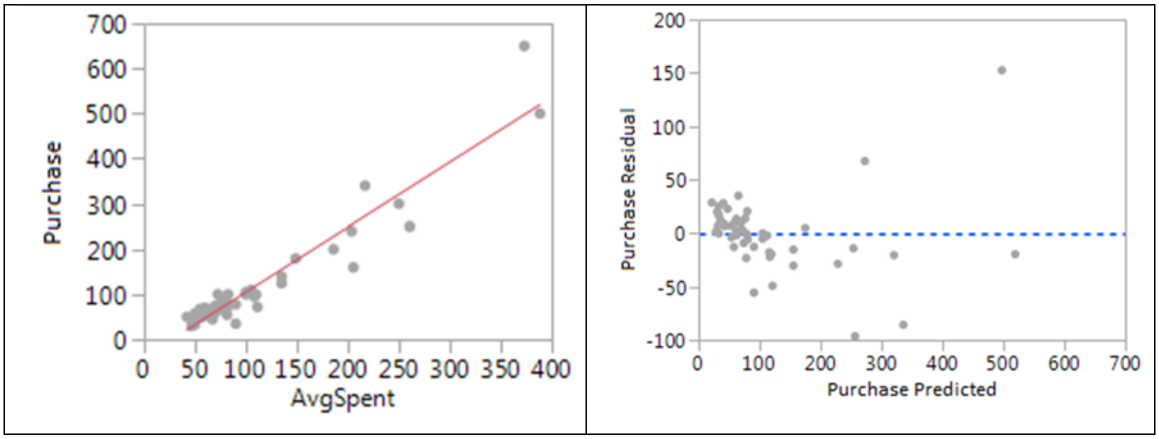
13. Find and interpret the value of R^2 for Model **2**. Compare this value to R^2 from the full multiple linear regression model in Part II and Model **1** here in Part III. Provide an argument that Model 2 is the best among these three models.

Part IV

We will now create a new explanatory variable AvgSpent = Total12/Number12 and use this variable in a simple linear regression model to predict the response variable, Purchase.

14. Out of the 56 observations in the sample, four of them will need to be eliminated from this simple linear regression model. Identify which observations (by Customer ID) must be removed. Describe how this change in the sample affects the definition of the population for inference.

15. Below is the scatterplot for the fitted simple linear regression model and its residual plot. The sample slope for the regression line is 1.44. Give the interpretation of this value in context.



16. Using the output in Problem **15**, describe the fit of this model and any potential problems with the standard model assumptions. For any issues you identify, suggest possible solutions which might mitigate or fix these issues.

Data Table

Cust. ID	Purchase	Months	Number12	Total12	Number24	Total24	Card
76	30	6	3	140	4	225	0
332	33	12	1	50	1	50	0
507	35	48	0	0	0	0	0
815	35	5	5	450	6	415	0
564	39	2	5	245	12	661	1
662	40	24	0	0	1	225	0
791	45	3	6	403	8	1138	0
761	48	6	3	155	4	262	0
699	50	12	1	42	7	290	0
249	50	5	2	100	8	700	1
612	50	8	3	144	4	202	0
132	50	1	10	562	13	595	1
394	50	2	3	166	4	308	0
298	50	4	4	228	4	228	0
166	50	5	5	322	7	717	1
965	55	13	0	0	6	1050	0
528	55	6	3	244	7	811	0
762	57	20	0	0	2	140	0
859	58	3	4	200	4	818	1
834	60	12	1	70	2	150	0
744	60	3	4	256	7	468	0
743	62	12	1	65	5	255	0
721	64	8	1	70	6	300	0
256	65	2	6	471	8	607	0
626	68	6	2	110	3	150	0
152	70	3	3	222	5	305	0
870	70	6	2	120	4	230	0
783	70	5	3	205	8	455	1
814	72	7	4	445	6	400	0
161	75	6	1	77	2	168	0
112	75	4	2	166	5	404	0
628	75	4	3	210	4	270	0
421	78	8	2	180	7	555	1
660	78	5	3	245	9	602	1
121	79	4	3	225	5	350	0
866	80	3	4	300	6	499	0
590	90	3	5	400	9	723	0
960	95	1	6	650	9	1006	1
514	98	6	2	215	3	333	0
790	100	12	1	100	2	200	0
686	100	2	1	110	4	400	1

442	100	3	3	217	6	605	0
240	100	3	4	330	8	660	1
495	105	2	4	400	7	560	0
733	110	3	4	420	6	570	0
219	125	3	2	270	5	590	1
622	140	6	3	405	6	775	0
908	160	2	2	411	8	706	0
738	180	1	5	744	10	945	1
770	200	1	3	558	4	755	1
419	240	4	4	815	10	1150	1
717	250	3	3	782	10	1500	1
510	300	12	1	250	4	401	0
484	340	1	5	1084	7	1162	1
149	500	4	2	777	3	905	1
351	650	1	4	1493	7	2050	1

A random sample of 56 customers from a large clothing retailer was selected. For each customer, the following variables were collected from store records.

Customer Number: Customer identification number for record keeping purposes.
Purchase: The net dollar amount spent by customers in their last purchase from this retailer.
Time: Number of months since last purchase.
Number12: Number of purchases in the last 12 months.
Amount12: Net total dollar amount of purchases in the last 12 months.
Number24: Number of purchases in the last 24 months.
Amount24: Net total dollar amount of purchases in the last 24 months.
Card: Whether or not the customer has a private label store credit card – 1 = Yes, 0 = No.

The table below contains the first five rows of the data table.

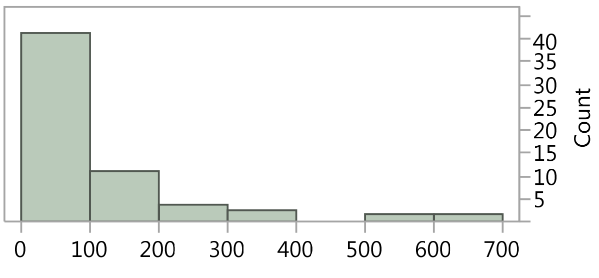
Customer Number	Purchase	Time	Number12	Amount12	Number24	Amount24	Card
76	30	6	3	140	4	225	0
332	33	12	1	50	1	50	0
507	35	48	0	0	0	0	0
815	35	5	5	450	6	415	0
564	39	2	5	245	12	661	1

The store manager is interested in predicting the amount a customer will spend on his or her next purchase using the explanatory variables: Time, Number12, Amount12, Number24, Amount24, and Card.

Part I

1. Below are graphical and numerical summaries of the response variable, Purchase. Describe the distribution of the variable. Based on your description, indicate any potential issues with using this response variable in a multiple linear regression analysis.

Purchase



Quantiles

100.0%	maximum	650
75.0%	quartile	100
50.0%	median	71
25.0%	quartile	50
0.0%	minimum	30

Summary Statistics

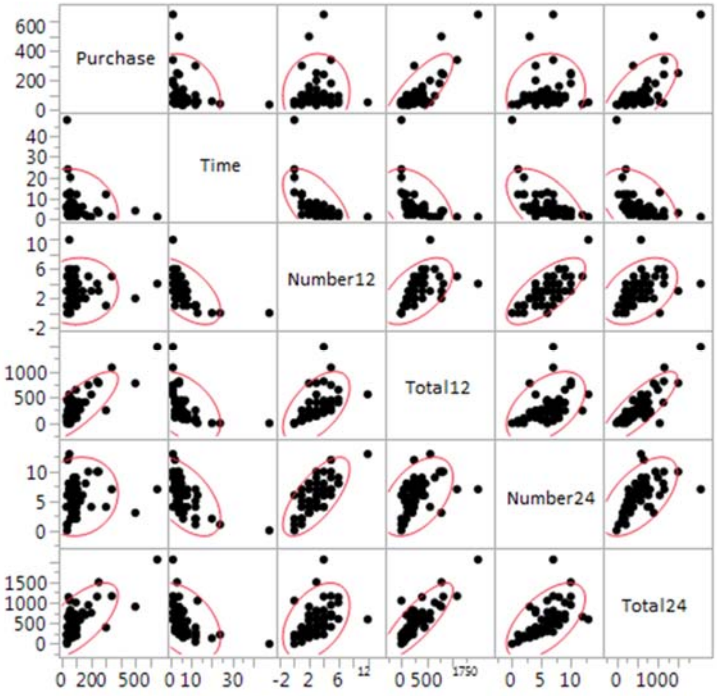
Mean	108.28571
Std Dev	112.18843
N	56

The distribution of the response variable Purchase is unimodal and skewed right, with a mean value of $108.29 and a standard deviation of $112.19. The five number summary is Min = $30, Q1 = $50, Median = $71, Q3 = $100 and Max = $650. There are several outliers in the distribution between $500 and $650.

The main potential issue with using this response variable in a multiple linear regression analysis will be the presence of the outliers. Depending on the explanatory variables, these observations may be difficult to model well and may contribute to larger standard errors.

2. Below is a correlation matrix and scatterplot of the quantitative variables in this analysis. Summarize the relationship between the response variable Purchase and the quantitative explanatory variables in the analysis. Which explanatory variables have the strongest and weakest relationship with the response variable, Purchase?

	Purchase	**Time**	**Number12**	**Total12**	**Number24**	**Total24**
Amount	1.0000					
Months	-0.2208	1.0000				
Number12	0.0516	-0.5838	1.0000			
Total12	0.8037	-0.4539	0.5559	1.0000		
Number24	0.1017	-0.5491	0.7100	0.4849	1.0000	
Total24	0.6773	-0.4324	0.4215	0.8275	0.5962	1.0000



The response variable Purchase has strong positive correlations with both Total12 and Total24. The scatterplot matrix shows a linear relationship between Purchase and these two variables, however, outliers might affect the strength of the relationship between Purchase and Total24 more than between Purchase and Total12.

The response variable Purchase has weak positive correlations with both Number12 and Number24. The scatterplot matrix shows a similar pattern for both relationships. In both cases, the outliers in the scatterplot tend to decrease the strength of the relationship between the two variables as measured by the correlation coefficient.

The response variable Purchase has a weak negative correlation with Time. The scatterplot matrix shows this correlation is likely influenced by one outlier in x = Time since last purchase.

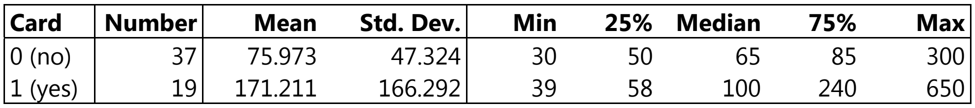
The variable with the strongest relationship with Purchase is Total12 and the weakest is Number12.

3. The output from Part I, Problem 2 also gives you a summary of the pairwise relationships among the quantitative explanatory variables. Summarize these relationships and indicate any potential issues with these relationships in the multiple linear regression analysis.

   All four explanatory variables have at least moderate correlations with each other. Based on the definitions of the explanatory variables, it is not surprising that the correlation between the variables Total12 and Total24 is the highest at $r = 0.8275$ with the correlation between the variables Number12 and Number24 also high at $r = 0.7100$. All pairwise correlations with the variable Time are negative. While this would generally indicate a decrease in the average value of the variables as the time since the last purchase increased, the scatterplots indicate this relationship is largely due to one outlier in the value of Time.

   The main potential issue with these relationship in the multiple linear regression analysis will be with multicollinearity – a high correlation between explanatory variables in the model. In this case, it may be difficult to separate the significance of the individual explanatory variables in the model. At the very least, we might expect to use only one of the Total variables and/or one of the Number variables in the final model.

4. Below are numerical summaries of the response variable, Purchase, between the two groups given by the explanatory variable Card. Summarize the differences in the distribution of purchases between the two groups. Does there seem to be a significant difference in the purchase amounts between the two groups? Explain your answer.



Card	Number	Mean	Std. Dev.	Min	25%	Median	75%	Max
0 (no)	37	75.973	47.324	30	50	65	85	300
1 (yes)	19	171.211	166.292	39	58	100	240	650

   The distribution of the response variable Purchase is very different between the two groups. For the customers without the store card, the distribution is right skewed, with this skewness primarily between the 75th percentile and the maximum observation. For the customers with the store card, the distribution is highly right skewed, with the skewness starting between the median and the 75th percentile. Between the two distributions, the lower tail of the distributions are similar ($30 to $39 for the minimums, $50 to $58 for Q1). However, the

difference between the Purchase amounts increase past the value of Q1, with $65 to $100 for the medians, $85 to $240 for Q3 and $300 to $650 for the maximum; indicating a higher center and higher variability for the distribution of Purchase for the customers with the store card. The means and standard deviations of the two distributions also indicate these differences, with a difference between the means of $75.97 to $171.21 and between the standard deviations of $47.32 to $166.29.

Given the information presented, the only statistical method available for assessing the significance of the difference between the two distributions is the t-test for the difference in the two population means. With such a large difference in standard deviations (ratio $= \frac{166.292}{47.324} = 3.51$), the t-test using Satterthwaite's approximation would be the more reasonable choice. The test statistic is

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{75.973 - 171.211}{\sqrt{\frac{(47.324)^2}{37} + \frac{(166.292)^2}{19}}} = -2.446$$

The smallest possible degrees of freedom using Satterthwaite's approximation is $n_2 - 1 = 18$. At this level, the test statistic will be statistically significant, indicating there is a statistically significant difference in the mean Purchase value between the two groups.

However, the skewed nature of the data distributions for both groups, the small sample sizes (particularly for the Card = Yes group), and the difference in the standard deviations of the two groups call into question the appropriateness of using a t-test for the difference in the two population means to assess the significance of the difference between the two groups. With access to the full data, using the non-parametric Wilcoxon Sum Rank Test might be a more appropriate choice.

Part II

For this part of the question, we will consider a multiple linear regression model to predict the variable Purchase from the all six explanatory variables.

5. Below is a table with the Type I Sums of Squares for the six explanatory variables in the model. Conduct a hypothesis test for the significance of the overall model in predicting the response variable, Purchase.

Variable	**Type I SS**
Time	33753
Number12	6278
Total12	565766
Number24	2987
Total24	74
Card	1679

There are several pieces of information given so far that must be used in order to create the ANOVA table for this multiple linear regression model.

Degrees of Freedom: With 6 explanatory variables in the model and a total of 56 observations, there will be 6 degrees of freedom for the Model, 55 for the Total and 55 – 6 = 49 for Error.

Sums of Squares: The Type I sums of squares given above can be summed to find the sums of squares for the Model. This is equal to 610537. The sums of squares for the Total can be found using information given in the distribution of the response variable in Part I, Problem 1. This will be equal to

$$SS_{Total} = s_y^2 * (n-1) = (112.18843)^2 * 55 = \ 692243.4$$

We can then find the sums of squares for Error by subtraction.

The ANOVA Table is:

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	**6**	**610537.0**	**101756.2**	**61.024**
Error	**49**	**81706.4**	**1667.48**	
Total	**55**	**692243.4**		

The F test statistic from the ANOVA table is very large, indicating the overall model is statistically significant in predicting the response variable, Purchase.

6. Below is the parameter estimates table for this multiple linear regression model. Use the information in the table to interpret the coefficients for the explanatory variables: Total12 and Total24. Then, conduct a hypothesis test for the significance of each variable in the model. In addition to the statistical result, explain why this finding might make sense given the context of the data.

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%	VIF
Intercept	111.56414	21.92715	5.09	<.0001*	67.499876	155.6284	.
Time	-1.345963	0.971053	-1.39	0.1720	-3.297368	0.6054422	1.6562382
Number12	-32.35354	5.18787	-6.24	<.0001*	-42.77895	-21.92812	3.0825433
Total12	0.4296834	0.041325	10.40	<.0001*	0.3466377	0.5127291	4.5409476
Number24	-5.173593	3.619661	-1.43	0.1593	-12.44757	2.1003879	3.2568561
Total24	0.0017562	0.03185	0.06	0.9562	-0.062248	0.0657608	4.8081163
Card[1]	14.624409	14.57577	1.00	0.3206	-7.3333487	21.95776	1.5994424

For the Total12 explanatory variable: For the multiple linear regression model including the explanatory variables Time, Number12, Number24, Total24 and Card and holding the values of these variables constant, a one dollar increase in the amount spent over the last 12 month period would be associated with a mean increase in the last purchase amount of approximately $0.43.

For the Total24 explanatory variable: For the multiple linear regression model including the explanatory variables Time, Number12, Time12, Number24, and Card and holding the values of these variables constant, a one dollar increase in the amount spent over the last 24 month period would be associated with a mean increase in the last purchase amount of approximately $0.0018.

With the other explanatory variables in the model, the variable Total12 is statistically significant with a p-value < 0.0001 while the variable Total24 is not statistically significant with a p-value of 0.9562.

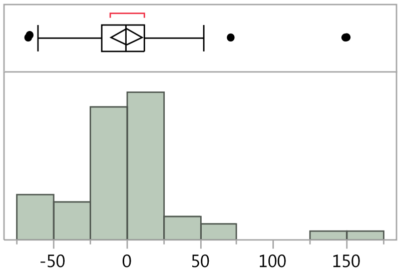
Given the context of the data, a one dollar increase in the variable Total12 would also mean a one dollar increase in the variable Total24. Given this relationship, it would make sense that only one of these two variables would be statistically significant in the model.

7. The table in Part II, Problem 6 also includes information about the VIF for each explanatory variable. Explain why the values for VIF are the highest for the Total12 and Total24 explanatory variables.

   Total12 and Total24 have the highest pairwise correlation between the explanatory variables. Total24 can also be calculated taking the value of Total12 and adding it to the value of the net total dollar amount in purchases for the second 12 month period.

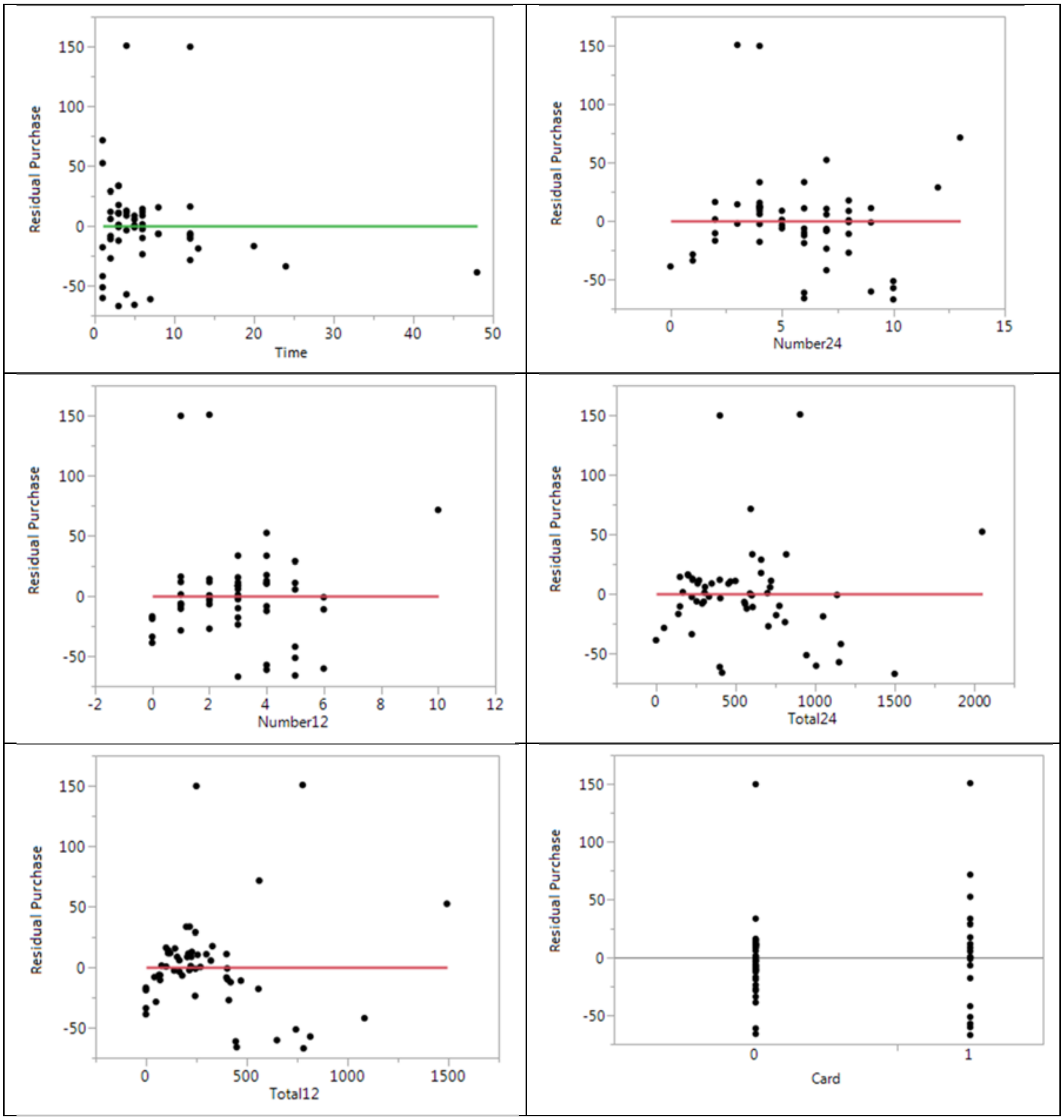
8. Below are graphical and numerical summaries for the distribution of residuals from the multiple linear regression model and residual plots for each explanatory variables. Use this information to check the assumptions of the model, making sure to indicate any potential issues.

Distribution of Residuals



```
Quantiles
 100.0%   maximum    150.62932217
  75.0%   quartile   11.662213361
  50.0%   median     -1.060655186
  25.0%   quartile   -17.64620711
   0.0%   minimum    -67.16053795
Summary Statistics
Std Dev               40.162288
```

Residual Plots



The model assumptions are that ε_i are i.i.d N(0, σ^2). The distribution of the residuals has several large outliers, indicating poor model fit for these observations. The remainder of the distribution of the residuals is unimodal, symmetric, and fairly bell-shaped. So the main issue here is with the outliers.

The residual plots also show these outliers, plus each explanatory variable has at least one outlier in the x-direction, indicating an observation with potentially large leverage and influence in the estimation of the model. While these plots don't seem to indicate any issues with the constant variance assumption, several of them contain a pattern in the residuals, indicating a potential need to add polynomial terms to the model.

9. The distribution of hat values (h_{ii}) and Cook's D values for this multiple linear regression model are given below. Use these values to indicate any potential observations with either high leverage, high influence or both.

   In looking at the distribution of the hat values, consider outliers and any observations greater than 2(k + 1)/n = 2(7)/56 = 0.25 or 3(k+1)/n = 3(7)/56 = 0.375. In both cases, there are two observations in the distribution with high leverage (obs. with hat values between 0.3 and 0.35 and between 0.4 and 0.45).

   In looking at the distribution of the Cook's D values, consider outliers and any observations with Cook's D value greater than approximately 1. No observations have a value greater than 0.7148, but there are 3 outliers in the distribution (obs. with Cook's D values between 0.6 and 0.7148) that are potentially influential.

Part III

For this part of the question, we will consider two multiple linear regression models to predict the response variable Purchase: Model 1 with Number12, Total12, and Card; and Model 2 with Number12 and Total12.

10. Using the sums of squares for Total from Part II, Problem 5, the ANOVA Table is

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	604150.6	201383.5	118.87
Error	52	88092.8	1694.09	
Total	55	692243.4		

   The large F-test statistic indicates the overall model is statistically significant in predicting the response variable Purchase.

11. The coefficient for Card in the model is $8.20. This means that given the same value of Number12 and Total12, the customer with the store card is expected to spend an average of $8.20 more on their last purchase than a customer without the store card. This coefficient is not statistically significant in the model with Number12 and Total12 as the p-value is 0.5470.

12. Comparing the response variable between the two groups in Part I, Problem 4 would indicate the response variable, Purchase, is different between the two groups. However, in the multiple linear regression model, the variable is not statistically significant. The multiple linear regression model also includes the variables Number12 and Total12. Once these variables are included in the model, the Card variable is not statistically significant. While we do not have access to the distributions of these variables between the two groups, given the relationship between Purchase and Total12, there is more than likely a difference in the distribution of Total12

between the two groups. Once Total12 is in the multiple linear regression model, the difference in the groups has been taken into account.

13. The value of the sums of squares for Model in Model 1 is 604150.6. Using the t-test statistic value from the coefficients table, we know the variable Card will decrease the sums of squares for Model (and increase the sums of squares for Error) for Model 2 by $t^2 = 0.61^2 = 0.37$. So, the value of R^2 from Model 2 is (604150.6 – 0.37)/692243.4 = 0.8727. This means that 87.27% of the variation in the response variable Purchase can be explained by the linear regression model with Number12 and Total12.

    The value of R^2 from the full model in Part II is **610537.0/692243.4 = 0.8820 and the value from Model 1 is 604150.6/692243.4 = 0.8727.**

    Model 2 is the best model of the three. The two variables are statistically significant, with the R^2 value of the model decreasing by slightly less than 1% compared to the full model, and only a very slight decrease from Model 1.

Part IV

We will now create a new explanatory variable AvgSpent = Total12/Number12 and use this variable in a simple linear regression model to predict the response variable, Purchase.

14. The 4 observations that must be eliminated from the analysis are the four customers who have not made a purchase in the past 12 months. These customers have a value of Number12 = 0.

    The sample is now just a sample of customers who have made a purchase in the last 12 months. This was not the original population the random sample was selected from. Now, we need to change our population for inference to be limited to just the customers who have made a purchase in the last 12 months.

15. The scatterplot indicates the model is a good fit to the data. However, both the scatterplot and the residual plot indicate a few problems with the model assumptions. There are outliers in the scatterplot and the residual plot. These are outliers in both x and y, indicating potential for high leverage and/or high influence. These points should be investigated to determine the effect on the fit of the model. Also, the residual plot indicate a pattern in the residuals, with positive residuals for lower predicted purchase values, and negative residuals for larger predicted purchase values. This indicates the potential for the addition of a polynomial term (square would be the most likely choice).

Part I

An experiment was conducted to study the effects of a virus on three plant genotypes (labeled 1, 2, and 3). Five plants of each genotype were grown in separate pots. All 15 pots were positioned in a single growth chamber using a completely randomized design. Two comparable leaves on each plant were marked for use in the experiment. One of the two marked leaves on each plant was randomly selected and injected with the virus. The other marked leaf on each plant was injected with a control substance. During the two weeks immediately following injections, virus-induced lesions formed on leaf surfaces. Two weeks after injections, a researcher measured the total lesion area separately for each of the two marked leaves on each of the 15 plants used in the experiment.

Let y_{ijk} be the total lesion area for genotype i ($i = 1, 2, 3$), plant j ($j = 1, \ldots, 5$), and leaf k, where $k = 1$ corresponds to the leaf injected with the control substance and $k = 2$ corresponds to the leaf injected with the virus. Let

$$\boldsymbol{y} = (y_{111}, y_{112}, y_{121}, y_{122}, y_{131}, y_{132}, y_{141}, y_{142}, y_{151}, y_{152}, y_{211}, y_{212}, \ldots, y_{341}, y_{342}, y_{351}, y_{352})'. \quad (1)$$

For $i = 1, 2, 3$ and $k = 1, 2$, assume $E(y_{ijk})$ is the same for all $j = 1, \ldots, 5$. Let $\mu_{ik} = E(y_{ijk})$ for all $i = 1, 2, 3$, $j = 1, \ldots, 5$, and $k = 1, 2$. Let

$$\boldsymbol{\beta} = (\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}, \mu_{31}, \mu_{32})'. \quad (2)$$

An appropriate model for the total lesion area data can be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e}, \quad (3)$$

where $\boldsymbol{y}$ and $\boldsymbol{\beta}$ are defined in (1) and (2), $\boldsymbol{X}$ and $\boldsymbol{Z}$ are matrices of constants, $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I})$ independent of $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma_e^2 \boldsymbol{I})$.

1. Provide the matrix $\boldsymbol{X}$.

2. Provide an appropriate matrix $\boldsymbol{Z}$.

Use the summary statistics on page 4 to complete problems **3** through **9**.

3. Estimate σ_u^2 .

4. Let $\bar{\mu}_{i\bullet} = (\mu_{i1} + \mu_{i2})/2$ for $i = 1, 2, 3$. Compute the value of the test statistic you would use to test $H_0 : \bar{\mu}_{1\bullet} = \bar{\mu}_{2\bullet} = \bar{\mu}_{3\bullet}$.

5. State the distribution of the test statistic in problem **4** under the assumption that the null hypothesis in problem **4** is true.

6. Let $\bar{\mu}_{\bullet k} = (\mu_{1k} + \mu_{2k} + \mu_{3k})/3$ for $k = 1, 2$. Compute the value of the test statistic you would use to test $H_0 : \bar{\mu}_{\bullet 1} = \bar{\mu}_{\bullet 2}$.

7. State the distribution of the test statistic in problem **6** under the assumption that the null hypothesis in problem **6** is true.

8. Some plant genotypes may be better than others at reducing the ability of the virus to spread from an initial infection point to other parts of the plant. To check for evidence of such differences among the three genotypes considered in this experiment, compute a test statistic that can be used to test

$$H_0 : \mu_{11} - \mu_{12} = \mu_{21} - \mu_{22} = \mu_{31} - \mu_{32} \, .$$

9. State the distribution of the test statistic in problem **8** under the assumption that the null hypothesis in problem **8** is true.

Part II

The researchers decided to investigate the ability of the virus to spread throughout an infected plant leaf. They constructed a new dataset using only the leaves from the experiment described in **Part I** that were directly injected with the virus. Each of the virus-injected leaves was partitioned into six adjacent sectors, numbered consecutively from 1 to 6. Sector 1 contains the base of each leaf where the virus was initially injected. Sector 6 contains the tip of the leaf and is the part of the leaf that is most distant from the initial injection site in sector 1. The following schematic diagram depicts the layout of sectors across each leaf.

injection at leaf base → | sector 1 | sector 2 | sector 3 | sector 4 | sector 5 | sector 6 | ← **leaf tip**

The sectors were chosen to be of equal area within each leaf. A researcher measured the total lesion area separately for each sector of each leaf. Let a_{ijs} be the total lesion area in sector s of the leaf injected with the virus on plant j of genotype i ($i = 1, 2, 3$, $j = 1, \ldots, 5$, $s = 1, \ldots, 6$). Page 5 contains `SAS` code and output associated with an analysis of the data $\{a_{ijs} : i = 1, 2, 3, \; j = 1, \ldots, 5, \; s = 1, \ldots, 6\}$.

10. Specify the model that was fit to the data using the `SAS` code on page 5. Define any new notation you introduce when specifying the model.

11. Explain what scientific question the researchers can address using the estimate labeled `Estimate 1` in the `SAS` output.

12. Compute a standard error associated with the estimate labeled `Estimate 1` in the `SAS` output.

13. Explain what scientific question the researchers can address using the estimate labeled `Estimate 2` in the `SAS` output.

14. Compute a standard error associated with the estimate labeled `Estimate 2` in the `SAS` output.

Part III

The researchers considered a third analysis that again involved only the leaves injected with the virus in the experiment described in **Part I**. Rather than measuring the total lesion area within each sector as in **Part II**, the researchers counted the total number of lesions for each leaf. They also measured the total area of each leaf. The resulting data set is provided in the `R` code and output that begins on page 6 and continues through page 9.

15. Compute BIC for each of the three models fit to the data in the `R` code.

16. Based on BIC, which of the three models is most appropriate?

17. Use the model 1 output to complete the blanks in the following sentence.

> The mean number of lesions per leaf for genotype 3 is estimated to be ______________ times the mean number of lesions per leaf for genotype 2. A confidence interval for this multiplicative factor that has coverage probability approximately equal to 95% is ______________ to ______________.

Summary Statistics for Part I

For all i, j, and k, let

$$y_{ij\bullet} = \frac{1}{2}\sum_{k=1}^{2} y_{ijk}, \quad y_{i\bullet k} = \frac{1}{5}\sum_{j=1}^{5} y_{ijk}, \quad y_{\bullet jk} = \frac{1}{3}\sum_{i=1}^{3} y_{ijk},$$

$$y_{i\bullet\bullet} = \frac{1}{10}\sum_{j=1}^{5}\sum_{k=1}^{2} y_{ijk}, \quad y_{\bullet j\bullet} = \frac{1}{6}\sum_{i=1}^{3}\sum_{k=1}^{2} y_{ijk}, \quad y_{\bullet\bullet k} = \frac{1}{15}\sum_{i=1}^{3}\sum_{j=1}^{5} y_{ijk},$$

$$\text{and} \quad y_{\bullet\bullet\bullet} = \frac{1}{30}\sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{2} y_{ijk}\ .$$

Using this notation, summary statistics are as follows:

$$\bar{y}_{11\bullet} = 30.8, \quad \bar{y}_{12\bullet} = 40.7, \quad \bar{y}_{21\bullet} = 42.7, \quad \bar{y}_{22\bullet} = 50.9, \quad \bar{y}_{31\bullet} = 41.6, \quad \bar{y}_{32\bullet} = 50.7,$$

$$\sum_{i=1}^{3}(\bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 = 77.0, \quad \sum_{j=1}^{5}(\bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet\bullet})^2 = 14.3, \quad \sum_{k=1}^{2}(\bar{y}_{\bullet\bullet k} - \bar{y}_{\bullet\bullet\bullet})^2 = 40.9,$$

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet})^2 = 176.8, \quad \sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{y}_{ij\bullet} - \bar{y}_{\bullet j\bullet})^2 = 518.8,$$

$$\sum_{i=1}^{3}\sum_{k=1}^{2}(\bar{y}_{i\bullet k} - \bar{y}_{i\bullet\bullet})^2 = 123.3, \quad \sum_{i=1}^{3}\sum_{k=1}^{2}(\bar{y}_{i\bullet k} - \bar{y}_{\bullet\bullet k})^2 = 154.7,$$

$$\sum_{j=1}^{5}\sum_{k=1}^{2}(\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet})^2 = 205.2, \quad \sum_{j=1}^{5}\sum_{k=1}^{2}(\bar{y}_{\bullet jk} - \bar{y}_{\bullet\bullet k})^2 = 29.5,$$

$$\sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{y}_{ij\bullet} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet j\bullet} + \bar{y}_{\bullet\bullet\bullet})^2 = 133.7,$$

$$\sum_{i=1}^{3}\sum_{k=1}^{2}(\bar{y}_{i\bullet k} - \bar{y}_{i\bullet\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet})^2 = 0.7,$$

$$\sum_{j=1}^{5}\sum_{k=1}^{2}(\bar{y}_{\bullet jk} - \bar{y}_{\bullet j\bullet} - \bar{y}_{\bullet\bullet k} + \bar{y}_{\bullet\bullet\bullet})^2 = 0.8,$$

$$\text{and} \sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{2}(\bar{y}_{ijk} - \bar{y}_{\bullet\bullet\bullet})^2 = 1749.3.$$

SAS Code and Output for Part II

```
proc mixed;
  class leaf genotype sector;
  model a = genotype sector genotype*sector;
  random leaf(genotype);
  repeated sector / subject = leaf(genotype) type=ar(1);
  estimate ’Estimate 1’ sector  1  0  0  0  0 -1
               genotype*sector  1  0  0  0  0 -1
                                0  0  0  0  0  0
                                0  0  0  0  0  0;
  estimate ’Estimate 2’ genotype  1  -1  0
                 genotype*sector  0  0  0  0  0  1
                                  0  0  0  0  0 -1
                                  0  0  0  0  0  0;
run;
```

```
       Covariance Parameter Estimates

Cov Parm           Subject           Estimate

leaf(genotype)                         0.4198
AR(1)              leaf(genotype)      0.7692
Residual                              0.09281


           Type 3 Tests of Fixed Effects

                    Num     Den
Effect               DF      DF    F Value    Pr > F

genotype              2      12       9.95    0.0028
sector                5      60     207.76    <.0001
genotype*sector      10      60      92.07    <.0001


               Estimates

                          Standard
Label         Estimate       Error

Estimate 1      6.5636      ??????
Estimate 2     -4.7178      ??????
```

R Code and Output for Part III

```
> ########################
> #                      #
> #       The Data       #
> #                      #
> ########################
>
> d3
   leaf genotype leafArea numberOfLesions
1     1        1     94.6              56
2     2        1    104.6              55
3     3        1     97.4              68
4     4        1    105.9              84
5     5        1    120.1              54
6     6        2    103.7              82
7     7        2    106.2              57
8     8        2    101.9              52
9     9        2    103.3              78
10   10        2    120.4              87
11   11        3    112.9              73
12   12        3    112.2              84
13   13        3     89.9              70
14   14        3    110.1             114
15   15        3    114.7              92
```

```
> ##############################
> #                            #
> #  Model 1 Code and Output   #
> #                            #
> ##############################
>
> o1 = glm(numberOfLesions ~ genotype, family=poisson(link = "log"), data = d3)
> summary(o1)

Call:
glm(formula = numberOfLesions ~ genotype, family = poisson(link = "log"),
data = d3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3912  -1.3571  -0.2808   1.0215   2.8066

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.14946    0.05617  73.879  < 2e-16 ***
genotype2    0.11603    0.07722   1.502    0.133
genotype3    0.31184    0.07392   4.219 2.46e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 56.77  on 14  degrees of freedom
Residual deviance: 38.10  on 12  degrees of freedom
AIC: 135.83

Number of Fisher Scoring iterations: 4

> vcov(o1)
             (Intercept)    genotype2    genotype3
(Intercept)  0.003154574 -0.003154574 -0.003154574
genotype2   -0.003154574  0.005963563  0.003154574
genotype3   -0.003154574  0.003154574  0.005464043
```

```
> ##############################
> #                            #
> #  Model 2 Code and Output   #
> #                            #
> ##############################
>
> o2=glm(numberOfLesions ~ leafArea + genotype, family=poisson(link = "log"),
+        data = d3)
> summary(o2)

Call:
glm(formula = numberOfLesions ~ leafArea + genotype,
    family = poisson(link = "log"), data = d3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1150  -1.3797  -0.4404   1.0338   2.6878

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.471772   0.383683   9.049  < 2e-16 ***
leafArea    0.006468   0.003613   1.790 0.073458 .
genotype2   0.100026   0.077697   1.287 0.197956
genotype3   0.289553   0.074957   3.863 0.000112 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 56.770  on 14  degrees of freedom
Residual deviance: 34.881  on 11  degrees of freedom
AIC: 134.61

Number of Fisher Scoring iterations: 4

> vcov(o2)
              (Intercept)      leafArea     genotype2     genotype3
(Intercept)  1.472123e-01 -1.371464e-03  9.300338e-05  1.562768e-03
leafArea    -1.371464e-03  1.305667e-05 -3.091772e-05 -4.491022e-05
genotype2    9.300338e-05 -3.091772e-05  6.036775e-03  3.260920e-03
genotype3    1.562768e-03 -4.491022e-05  3.260920e-03  5.618518e-03
```

```
> ##############################
> #                            #
> #  Model 3 Code and Output   #
> #                            #
> ##############################
>
> o3=glm(numberOfLesions ~ I(log(leafArea)) + genotype,
+        family=poisson(link = "log"), data = d3)
> summary(o3)

Call:
glm(formula = numberOfLesions ~ I(log(leafArea)) + genotype,
    family = poisson(link = "log"), data = d3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1151  -1.3943  -0.4132   1.0417   2.6637

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       0.88025    1.77437   0.496 0.619832
I(log(leafArea))  0.70331    0.38133   1.844 0.065130 .
genotype2         0.09855    0.07775   1.267 0.204992
genotype3         0.28912    0.07495   3.857 0.000115 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 56.77  on 14  degrees of freedom
Residual deviance: 34.67  on 11  degrees of freedom
AIC: 134.4

Number of Fisher Scoring iterations: 4

> vcov(o3)
                 (Intercept) I(log(leafArea))    genotype2    genotype3
(Intercept)       3.14839935     -0.676280149  0.012901073  0.018851875
I(log(leafArea)) -0.67628015      0.145411525 -0.003452232 -0.004731754
genotype2         0.01290107     -0.003452232  0.006045523  0.003266911
genotype3         0.01885187     -0.004731754  0.003266911  0.005618016
```

Part I

1. $\boldsymbol{X} = \boldsymbol{I}_{3\times 3} \otimes \mathbf{1}_{5\times 1} \otimes \boldsymbol{I}_{2\times 2}$

2. The most straightforward answer is $\boldsymbol{Z} = \boldsymbol{I}_{15\times 15} \otimes \mathbf{1}_{2\times 1}$. Any matrix that has the same 15 columns as $\boldsymbol{Z}$ (in any order) is acceptable.

3. This is a split-plot experiment with genotype as the whole-plot factor, plant nested within genotype as the whole-plot experimental unit, injection as the split-plot factor, and leaf as the split-plot experimental unit. Let u_{ij} be the random effect associated with the jth plant of genotype i. With the choice of $\boldsymbol{Z}$ given in the solution to problem **2**, we have

$$\boldsymbol{u} = (u_{11}, u_{12}, u_{13}, u_{14}, u_{15}, u_{21}, u_{22}, u_{23}, u_{24}, u_{25}, u_{31}, u_{32}, u_{33}, u_{34}, u_{35})'.$$

The standard ANOVA table for such a split-plot experiment is given by

Source	DF	Sum of Squares
genotype	$3-1$	$5 * 2 * \sum_{i=1}^{3}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 = 5 * 2 * 77.0$
plant(genotype)	$(5-1)3$	$2 * \sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot})^2 = 2 * 176.8$
injection	$2-1$	$3 * 5 * \sum_{k=1}^{2}(\bar{y}_{\cdot\cdot k} - \bar{y}_{\cdot\cdot\cdot})^2 = 3 * 5 * 40.9$
genotype × injection	$(3-1)(2-1)$	$5 * \sum_{i=1}^{3}\sum_{k=1}^{2}(\bar{y}_{i\cdot k} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot k} + \bar{y}_{\cdot\cdot\cdot})^2 = 5 * 0.7$
error	(2-1)(5-1)3	$1749.3 -$ (sum of lines above) $= 8.7$
corrected total	$30-1$	$\sum_{i=1}^{3}\sum_{j=1}^{5}\sum_{k=1}^{2}(\bar{y}_{ijk} - \bar{y}_{\cdot\cdot\cdot})^2 = 1749.3$

Note that the error line can alternatively be labeled as injection × plant(genotype), which gives some insight into the expression for the error degrees of freedom presented in the table. Of course, the error degrees of freedom can also be obtained by subtracting the sum of the other degrees of freedom from the corrected total of 29.

The expected mean square for plant(genotype) is

$$
\begin{aligned}
E\left\{\frac{2}{(5-1)3}\sum_{i=1}^{3}\sum_{j=1}^{5}(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot})^2\right\} &= \frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{5}E(\bar{y}_{ij\cdot}-\bar{y}_{i\cdot\cdot})^2\\
&= \frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{5}E(\bar{u}_{ij}-\bar{u}_{i\cdot}+\bar{e}_{ij\cdot}-\bar{e}_{i\cdot\cdot})^2\\
&= \frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{5}\left\{E(\bar{u}_{ij}-\bar{u}_{i\cdot})^2+E(\bar{e}_{ij\cdot}-\bar{e}_{i\cdot\cdot})^2\right\}\\
&= \frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{5}\left\{\mathrm{Var}(\bar{u}_{ij}-\bar{u}_{i\cdot})+\mathrm{Var}(\bar{e}_{ij\cdot}-\bar{e}_{i\cdot\cdot})\right\}\\
&= \frac{1}{6}\sum_{i=1}^{3}\sum_{j=1}^{5}(\sigma_u^2+\sigma_u^2/5-2\sigma_u^2/5+\sigma_e^2/2+\sigma_e^2/10-2\sigma_e^2/10)\\
&= \frac{15}{6}\left(\frac{4}{5}\sigma_u^2+\frac{4}{10}\sigma_e^2\right)\\
&= 2\sigma_u^2+\sigma_e^2.
\end{aligned}
$$

Furthermore, we know the expected error mean square is σ_e^2. Thus, a method of moments estimate of σ_u^2 (that matches the REML estimate in this case) is given by

$$\{2*176.8/12-8.7/12\}/2\approx 14.4.$$

4. This is a test for genotype main effects. From the ANOVA table generated in the solution to problem **3**, we have

$$F=\frac{5*2*77.0/2}{2*176.8/12}\approx 13.1.$$

Note that the error term for testing genotype main effects comes from plant(genotype) because plants nested in genotypes are the whole-plot experimental units.

5. The null distribution of the test statistic is F with 2 numerator degrees of freedom and 12 denominator degrees of freedom.

6. This is a test for injection main effects. From the ANOVA table generated in the solution to problem **3**, we have

$$F=\frac{3*5*40.9/1}{8.7/12}\approx 846.2.$$

7. The null distribution of the test statistic is F with 1 numerator degrees of freedom and 12 denominator degrees of freedom.

8. This is a test for genotype × injection interaction effects. From the ANOVA table generated in the solution to problem **3**, we have

$$F = \frac{5 * 0.7/2}{8.7/12} \approx 2.4.$$

9. The null distribution of the test statistic is F with 2 numerator degrees of freedom and 12 denominator degrees of freedom.

Part II

10. For $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$, let $\boldsymbol{\epsilon}_{ij} = (\epsilon_{ij1}, \ldots, \epsilon_{ij6})'$. Let $\sigma_\epsilon^2 > 0$ and $\rho \in (-1, 1)$ be unknown parameters, and let

$$\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} = \sigma_\epsilon^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 & \rho^5 \\ \rho & 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho^2 & \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^3 & \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^5 & \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}.$$

Suppose $\boldsymbol{\epsilon}_{11}, \ldots, \boldsymbol{\epsilon}_{35} \overset{i.i.d.}{\sim} N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. For $i = 1, 2, 3$, $j = 1, \ldots, 5$, and $s = 1, \ldots, 6$, let

$$a_{ijs} = \theta_{is} + \ell_{ij} + \epsilon_{ijs},$$

where θ_{is} is an unknown parameter ($i = 1, 2, 3$ and $s = 1, \ldots, 6$), the random leaf effects ℓ_{ij} for $i = 1, 2, 3$ and $j = 1, \ldots, 5$ are independent and identically distributed as $N(0, \sigma_\ell^2)$ for some unknown positive variance component σ_ℓ^2, and all the leaf random effects are independent of $\boldsymbol{\epsilon}_{11}, \ldots, \boldsymbol{\epsilon}_{35}$.

11. For genotype 1, what is the mean total lesion area in sector 1 minus the mean total lesion area is sector 6? A value near 0 may be evidence that the virus is able to spread from the injection site to leaf tips in genotype 1. A large value could be evidence that the virus does not move so easily from the injection site to the leaf tip in genotype 1 leaves.

12.

$$\begin{aligned} \text{Var}(\bar{a}_{1\cdot 1} - \bar{a}_{1\cdot 6}) &= \text{Var}(\bar{\epsilon}_{1\cdot 1} - \bar{\epsilon}_{1\cdot 6}) \\ &= \frac{1}{5}\text{Var}(\epsilon_{111} - \epsilon_{116}) \\ &= \frac{1}{5}[2\sigma_\epsilon^2 - 2\text{Cov}(\epsilon_{111}, \epsilon_{116})] \\ &= \frac{1}{5}(2\sigma_\epsilon^2 - 2\sigma_\epsilon^2\rho^5) \\ &= \frac{2\sigma_\epsilon^2(1 - \rho^5)}{5} \end{aligned}$$

Thus, the standard error is $\sqrt{2(0.09281)(1-0.7692^5)/5} \approx 0.1647$.

13. What is the difference between the mean total lesion area in the tips of genotype 1 leaves and the mean total lesion area in the tips of genotype 2 leaves?

14.

$$\begin{aligned}\text{Var}(\bar{a}_{1\cdot 6}-\bar{a}_{2\cdot 6}) &= \text{Var}(\bar{\ell}_{1\cdot}-\bar{\ell}_{2\cdot}+\bar{\epsilon}_{1\cdot 6}-\bar{\epsilon}_{2\cdot 6}) \\ &= \frac{2}{5}(\sigma_\ell^2+\sigma_\epsilon^2)\end{aligned}$$

Thus, the standard error is $\sqrt{2(0.4198+0.09281)/5} \approx 0.4528$.

Part III

15. $\text{BIC} = \text{AIC} - 2k + k\log(n)$, where k is the dimension of the model's parameter space and n is the number of observations.

Model	BIC
1	$135.83 - 2(3) + 3\log(15) \approx 137.95$
2	$134.61 - 2(4) + 4\log(15) \approx 137.44$
3	$134.40 - 2(4) + 4\log(15) \approx 137.23$

16. Model 3 would be deemed most appropriate according to BIC because Model 3 has the lowest BIC value.

17. The multiplicative factor that fills in the first blank is $\exp(0.31184-0.11603) \approx 1.22$. A check for overdispersion suggests a potential problem with the Model 1 fit. The residual deviance is 38.10. Compared to a central chi-square distribution with $15-3=12$ degrees of freedom, this value is unusually large (right tail area 0.00015). Thus, we should adjust for potential overdispersion by multiplying the variance of our coefficient vector estimator by $38.10/12$ and using t quantiles when contructing our confidence interval. The standard error associated with the estimate $0.31184-0.11603$ is

$$\sqrt{(38.10/12)(0.005464043+0.005963563-2*0.003154574)} \approx 0.1275.$$

The confidence interval for the multiplicative factor is provided in the expression below, where 2.18 is (approximately) the 0.975 quantile of a t-distribution with 12 degrees of freedom.

$\exp(0.31184-0.11603-2.18*0.1275) \approx 0.92$ to $\exp(0.31184-0.11603+2.18*0.1275) \approx 1.60$

A mathematical graph or network consists of a set of nodes and a set of edges that join some or all pairs of nodes. Nodes are often associated with objects such as people, countries, or genes. Edges are associated with some type of relation between nodes such as co-authorship of papers, the existence of trade agreements, or disease status. Probabilistic models for graphs usually assign binary random variables to potential edges, and specify probabilities that edges are present in a realization of the model. To set up the problem in this manner consider a graph with n fixed nodes, and let $Y_1, \ldots, Y_k$ be binary random variables corresponding to the $k = n(n-1)/2$ potential edges such that Y_i represents the potential edge joining nodes $s_i \in \{1, \ldots, n\}$ and $u_i \in \{1 \ldots, n\}$, $s_i \neq u_i$. Define, for $i = 1, \ldots, k$,

$$Y_i = \begin{cases} 1 & \text{if edge } i \text{ is realized} \\ 0 & \text{otherwise.} \end{cases}$$

A random graph model is completed by specifying or building a joint distribution for these random variables.

The simplest random graph model is known as the Erdos-Renyi graph model. Under this model, $Y_1, \ldots, Y_k$ are assumed to be independent and to have the same probability of being realized, so that each variable has the probability mass function, for $0 < p < 1$,

$$f(y_i|p) = p^{y_i}\,(1-p)^{1-y_i};\,,\;\; y_i = 0, 1.$$

The joint distribution of $Y_1, \ldots, Y_k$ is then

$$f(\boldsymbol{y}|p) = \prod_{i=1} n f(y_i|p),$$

and has support on the $k-$fold Cartesian product of $\{0, 1\}$.

An extension of an Erdos-Renyi graph model results in what are called “block” models. In block models, the nodes are divided into mutually exclusive sets based on some characteristics. The simplest way to conceptualize this is having nodes of different colors. Blocks are then defined relative to the types of nodes joined by potential edges, such as red-red edges, blue-blue edges, and red-blue edges, and each block has a constant probability for potential edges within it, combined with independence. So block models correspond to sets of overlaid Erdos-Renyi models (one for each block). Other models may use auxiliary information about nodes to define continuous covariates on which the probability of edge realization depends. An example might be using distance among nodes that represent cities to model the probability of edges that correspond to direct (non-stop) airflights. An assumption of independence among random variables that correspond to potential edges is usually maintained in these models. Note that block models and models using covariates are ANOVA and regression models for binary response variables.

More complicated models are often constructed to be what are called Exponential Random Graph (ERG) models. ERG models are constructed by counting features of graph topology (e.g., the number of edges and the number of triangles) to be used in the role of sufficient statistics in a joint distribution with an overall Gibbsian form. The random variables corresponding to edges are no longer independent in ERG models, but it can be difficult to quantify the structure and degree of dependencies that may exist. Finally, Local Structure Graph (LSG) models use a binary Markov random field to model edge probabilities through specification of full conditional distributions and also result in a Gibbsian form for joint distributions of edge realization. LSG models allow control over the structure and strength of dependencies in edge realizations, but require the definition of neighborhoods for potential edges. More information on ERG and LSG models will be presented when it is needed. Right now, the important point is that these models do not assume independence among random variables that correspond to potential edges.

It is sometimes claimed that before one can legitimately promote a more complex model as an alternative for a given situation, one must first be able to reject the simple Erdos-Renyi model as appropriate to represent that situation.

ANSWER QUESTIONS 1, 2, 3, 4 AND 5 NOW

(Questions begin on page 5.)

Now consider models for random graphs that do not assume that the realization of edges are independent, that is, do not assume the random variables Y_i; $i = 1, \ldots, k$ are independent of one another. As a concrete example of a situation in which we might want to consider the use of such models, suppose that we would like to fit a random graph model to data on international trade relations. Such data are presented for 10 countries in graphical form in Figure 1 (on page 7). The countries form nodes of the graph, and are assigned a spatial position based on two indices, political gradient and size of economy, both scaled to the range (0, 10). Size of economy is straightforward. Political gradient represents the extent to which a country has a political system that differs from that of the United States. Thus, two countries that differ from the United States by the same amount are very similar in terms of political system, and political gradient can be considered a measure of political similarity between two countries. The United States is not one of the countries depicted in Figure 1, but would lie at coordinate $(10, 0)$ if it were included on the graph. It is natural to suspect that the probability of edge formation in the situation of Figure 1 might be a function of distance between nodes in terms of the constructs of political similarity and size of economy. Political scientists also have theorized, however, that in terms of trade, countries tend to behave in the same way toward other countries that are roughly the same displacement from them in the dimensions of Figure 1. The implications of these two theories are that, if a given country has two other countries located close to each other on Figure 1 (about the same displacement from the focal country), it should tend to either form trade relations with both, or neither, with both being more likely if the distances are smaller and neither being more likely if the distances are greater.

To model the graph of Figure 1 with this theoretical underpinning in mind, the "locations" of both realized and potential edges were defined as the center of the line segment described by the edge. These were considered locations in a random field and were then attached to a total of $k = 45$ random variables defined to correspond to possible edges between the 10 nodes of the original graph in Figure 1. Neighborhoods for these random field locations were then defined using a prescribed distance – locations within a certain radius of a focal location were considered to be neighbors of that location. The effect of this is that countries in the original graph that differ from a focal country by about the same displacement will have edge "locations" near each other and thus in the same neighborhood. A Local Structure Graph model was formulated by specifying full conditional probability mass functions, for $0 < p_i < 1$ and $i = 1, \ldots, k$,

$$f(y_i|p_i) = p_i^{y_i}\,(1-p_i)^{1-y_i}; \quad y_i = 0, 1 \tag{1}$$

where, for $0 < \kappa_i < 1$ and $-\infty < \eta_{i,j} < \infty$,

$$\log\left(\frac{p_i}{1-p_i}\right) = \log\left(\frac{\kappa_i}{1-\kappa_i}\right) + \sum_{j=1}^{k} \eta_{i,j}(y_j - \kappa_j), \tag{2}$$

subject to $\eta_{i,j} = 0$ unless $j \neq i$ and the location of Y_j is in the neighborhood of Y_i, in which case $\eta_{i,j} = \eta$, a constant. The effect of distance (in terms of the axes of Figure 1) on the marginal probability of realization for each edge was incorporated as,

$$\log\left(\frac{\kappa_i}{1-\kappa_i}\right) = \beta_0 + \beta_1 d(s_i, u_i), \tag{3}$$

where $d(s_i, u_i)$ is the distance between nodes of the original graph in Figure 1 (s_i and u_i) that are joined by edge i, and $-\infty < \beta_0 < \infty$ and $-\infty < \beta_1 < \infty$ are regression parameters.
Note that there are two graphs, and hence two definitions of nodes and two definitions of edges, that are involved in this model.

1. Original Graph. This is a random graph.

Nodes: Countries

Edges: Trade relations between countries

2. Random Field Graph. This is a fixed graph.

Nodes: Locations of centerpoints of original graph potential edges

Edges: Connect neighboring locations in the Random Field Graph

Recall from class that under the model of expressions (1) through (3) the joint probability mass function for $Y_1, \ldots, Y_k$ has the form, for a complicated function Q that we do not need to bother with here,

$$f(\boldsymbol{y}|\beta_0, \beta_1, \eta) = \frac{\exp[Q(\boldsymbol{y}, \beta_0, \beta_1, \eta)]}{\sum\limits_{\boldsymbol{y}\in\Omega} \exp[Q(\boldsymbol{y}, \beta_0, \beta_1, \eta)]}, \tag{4}$$

where Ω is the $k-$fold Cartesian product of $\{0, 1\}$. The important point relative to expression (4) is that the joint distribution is computationally prohibitive due to the number of terms in the sum in the denominator.

ANSWER QUESTIONS 6, 7, 8, and 9 NOW

(Questions on page 6.)

Return to the issue that this prelim question started with – could we reject a simple Erdos-Renyi model as being an adequate representation of the data of Figure 1? We certainly do not run into any issue in estimation of an Erdos-Renyi model. The observed graph of Figure 1 has a fixed number of $n = 10$ nodes, and 19 realized edges, so a maximum likelihood estimate of the constant Erdos-Renyi probability of edge realization is $\hat{p} = 19/45 = 0.422$. A single realization of this fitted model is presented in Figure 2 (page 8) for the same 10 nodes as Figure 1. In this particular realization there are 21 edges. We could easily get similar realizations from the posterior predictive distribution in a Bayesian procedure. This, and other realizations could be visually compared to the observed graph of Figure 1, and we might like to think we see a difference in the structure of these two pictures, but for scientific inference we need more than this. In fact, under an Erdos-Renyi model, the graph of Figure 1 and the graph of Figure 2 have equal probability. So our need becomes to either (1) reject Erdos-Renyi outright to motivate consideration of our more complex LSG model alternative or (2) demonstrate that our alternative results in a more pleasing (or better) representation of the data than does the Erdos-Renyi model.

ANSWER QUESTION 10 NOW

(Question on page 6.)

Questions

1. Under the concept that one must reject an Erdos-Renyi model as appropriate for a given situation before considering any other models, what is meant by the "situation" is critical, and depends on the reason one is conducting an analysis of a particular problem. If there is a feature of the graph that is connected with the underlying scientific problem, then one can often use that feature as a test statistic. For example, in social science the concept of "transitivity" is related to the number of triangles that are contained in a graph. Suppose that, in a problem for which the focus is the concept of transitivity, we have an observed graph that contains n nodes, e edges, and t triangles. Outline in algorithmic form a procedure to test that our observed graph can be represented by an Erdos-Renyi graph model. Assume that we have a computational function called `notris` that can take a list of nodes and edges and compute the number of triangles.

2. Consider block models or models that take edge probability to be a function of some continuous covariates. Let Y_i; $i = 1, \ldots, k$ denote binary random variables connected with the realization of edge i, where $k = n(n-1)/2$, and let $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,p})$; $i = 1, \ldots, k$ denote covariates associated with these edges (these may be either grouping variables or continuous covariates). Assume that the edges are arbitrarily indexed. Assuming independence among the Y_i, formulate a model that would be appropriate for either a situation with blocks of edge types, or edge probabilities that depend on continuous covariates. Clearly indicate any modeling choices that must be made in this development, and recall that a model must result in a joint probability distribution for all of the random variables involved in the problem.

3. Discuss whether or not you could use your algorithm from Question 1 (with some test statistic other than the number of triangles) to reject an Erdos-Renyi model for a situation in which you think you might like to use your model from Question 2. If you believe you could do so, be certain to identify an appropriate test statistic (which cannot depend on fitting the model of Question 2). If you believe you could not do so, identify the source of the reason (the problem). Note that the **only** change to your algorithm of Question 1 allowed is the definition of the test statistic. **Do not** (yet) provide an alternative algorithm or procedure that might address this question.
Hint: Either position could lead to an adequate answer in this question. The question deals with whether or not we can consider Erdos-Renyi strictly in terms of model adequacy rather than model selection – that is the effect of the stipulation that the only thing you can change in the algorithm of Question 1 is the test statistic.

4. **Now** present a procedure that could be used to determine whether your model from Question 2 should be preferred to a simpler Erdos-Renyi model for a particular data set. That is, now approach the problem from the viewpoint of model selection. Note that there may be a number of good procedures available for this. Don't try to present all of them, just one you would use. For this question your procedure may well depend on fitting the model of Question 2 to one or more sets of values.

5. The concepts that underlie questions 3 and 4 present an issue related to the topics of model adequacy and assessment. This issue is related to viewpoints that take model assessment to be primarily a question of adequate fit, versus primarily a question of model selection. What is the danger in viewing model assessment as purely a question of model selection? What is the danger in viewing model assessment as purely a question of adequacy of fit?

6. If you were given the numerical version of the data in Figure 1, how would you estimate the parameters of the model of expressions (1) through (4) in a non-Bayesian analysis? In particular, give the form of an objective function that you would use for parameter estimation.

7. If the parameter η in this LSG model is 0, then the random variables $Y_1, \ldots, Y_k$ are independent, the full conditional distributions of expressions (1) through (3) are also marginal distributions, and the model reduces to the situation considered in Question 2. Still in a non-Bayesian framework, how would you assess the evidence offered by the data that $\eta = 0$ (or $\eta \neq 0$)? Be reasonably specific, but you may refer to standard procedures by name – for example, if you would use a score test you may simply say that (for a specified hypothesis) BUT the key would be how you would compute the inverse information evaluated at the hypothesized parameter value, so you would need to address that. Again, names of procedures are sufficient here, no formulae are required (but okay if you want).

8. For whatever procedure you suggested in Question 7, identify the most difficult aspect or aspects of using that procedure. These may be computational difficulties, or they may be difficulties in motivating or justifying use of the procedure.

9. Whatever difficulty or difficulties you may have identified for a non-Bayesian analysis in Question 8 could be overcome by basing inference on a posterior distribution in a Bayesian analysis. What would be the greatest hurdle to be overcome to enact a Bayesian strategy with this model?

10. Suggest an approach (perhaps in outline form, but **not** in algorithmic form) to either reject (or fail to reject) an Erdos-Renyi model as an adequate representation of the data of Figure 1 OR to demonstrate that the more complex LSG model provides a sufficiently better representation so as to be preferred to the Erdos-Renyi model.
Hint: Consideration of the following may help you formulate an answer:

- Is there a feature of the data readily available that captures the scientific distinction between an Erdos-Renyi model and the LSG model?
- Consider that there may be models that are "in between" the Erdos-Renyi model and the LSG model.
- Review your answers to Questions 4 and 7.

Figures

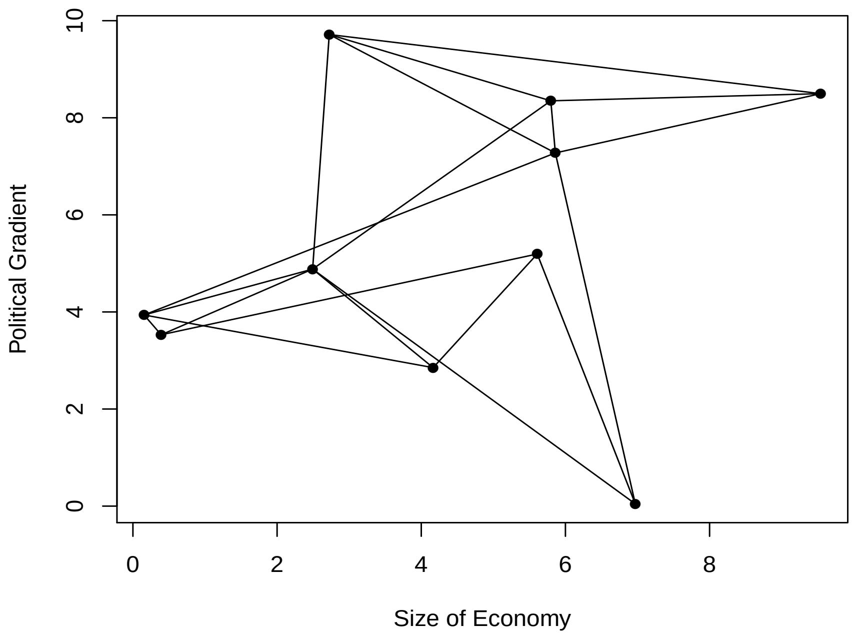


Figure 1: A graph of international relations based on political similarity and size of economy of 10 countries. Edges represent trade relations. There are 19 edges in this graph.

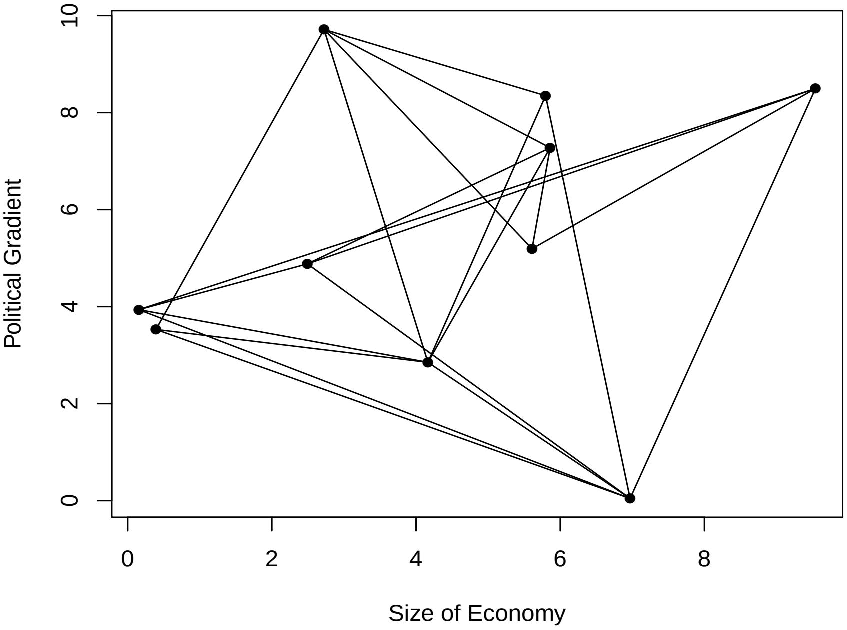


Figure 2: Realization of an Erdos-Renyi graph fit to the data of Figure 1. There are 21 edges in this graph.

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

Question 1. To set up the problem, let $Y_1, \ldots, Y_k$ be binary random variables corresponding to the $k = n(n-1)/2$ potential edges such that

$$Y_i = \begin{cases} 1 & \text{if edge } i \text{ is realized} \\ 0 & \text{otherwise.} \end{cases}$$

Assume that the Y_i are independent and identically distributed with binary probability mass functions, for $0 < p < 1$,

$$f(y_i|p) = p^{y_i}\,(1-p)^{1-y_i}; \quad y_i = 0, 1.$$

The random variables Y_i are assigned to join nodes s_i and u_i in the observed graph, where $s_i, u_i \in \{1, \ldots, n\}$ and pairs (s_i, u_i) are mutually exclusive and exhaustive of the list of unique pairs. For an observed graph with n nodes, e edges, and t triangles, a Monte Carlo (or simulation-based) procedure to test the adequacy of an Erdos-Renyi model as a probabilistic representation is as follows.

(a) Estimate the constant probability of edge realization as

$$\hat{p} = \frac{e}{n(n-1)/2} = \frac{2e}{n(n-1)}.$$

(b) For $m = 1, \ldots, M$,

i. Simulate a realization for k independent binary random variables with probability $\hat{p}$. Denote this realization as $\boldsymbol{y}^*_m = (y^*_1, \ldots, y^*_k)_m$. Using the same association between edge and node indices as for the observed graph, prepare a list of nodes and edges in whaterver format is expected by the function `notris`.

ii. Using the function `notris`, compute the number of triangles in the graph represented by $\boldsymbol{y}^*_m$, and call this value t^*_m.

(c) Using $I(A)$ to denote the identity function that assumes the value of 1 if A is true and 0 otherwise, calculate a Monte Carlo $p-$value, p_M, for the hypothesis that the Erdos-Renyi model is adequate to represent the transitivity of the observed graph as

$$\begin{aligned} p_{M,r} &= \sum_{m=1}^{M} I(t^*_m \leq t) \\ p_{M,\ell} &= \sum_{m=1}^{M} I(t \leq t^*_m) \\ p_M &= \min\{p_{M,\ell}, p_{M,r}\}. \end{aligned} \tag{1}$$

Question 2. An appropriate model for this question has the form of a basic generalized linear model (glm) with binary random component. Take the random variables $Y_1, \ldots, Y_k$ to be defined as in Question 1. Assume that these variables are independent with probability mass functions, for $0 < \theta_i < 1$ and $i = 1, \ldots, k$,

$$f(y_i|\theta_i) = \theta_i^{y_i}\,(1-\theta_i)^{1-y_i}; \quad y_i = 0, 1. \tag{2}$$

Now model the parameters as, for $i = 1, \ldots, k$ and $-\infty < \beta_j < \infty; \quad j = 1, \ldots, p$,

$$g(\theta_i) = \sum_{j=1}^{p} \beta_j\, x_{i,j}, \tag{3}$$

where $g(\cdot)$ is a known link function with domain $(0, 1)$ and range the entire real line, such as a logit or complementary log-log link. The most common choice will probably be a logit link,

$$g(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right).$$

The link function is a choice that must be made in formulation of this model, and is more important for models with continuous covariates than for block models in which the $x_{i,j}$ are indicators of block membership. In this latter case, equivalent results will occur for any monotone link function. The joint probability mass function for $Y_1, \ldots, Y_k$ is then

$$f(\boldsymbol{y}|\beta_1, \ldots, \beta_p) = \prod_{i=1}^{k} f(y_i|\theta_i), \tag{4}$$

where $f(y_i|\theta_i)$ is given in (1) and θ_i is given in (2).

Question 3. Either position could lead to an adequate answer here, depending on the motivation and justification. The most straightforward position would be that use of the Question 1 algorithm in this situation would be difficult because the model of Question 2, in and of itself, does not suggest an unambiguous test statistic. That is, the model does not *directly* indicate a feature of the graph that is connected with substantive reason for conducting an analysis which, as indicated in the preamble to Question 1 is a critical factor underlying the idea for testing the adequacy of an Erdos-Renyi model in isolation from any potential alternative. On the other hand, the purpose of the model is to relate the probabilities of edge realization to values of a covariate. With binary response variables, it is not obvious how one could quantify this idea without specification of the mathematical form of that relation (that is, without the use of an alternative model), but one could make some attempts. For covariates that indicate block membership, the difference in the probabilities of edge realization between two blocks would work, and could perhaps be extended to more than two blocks as the sum of the squared probabilities for blocks. For covariates that are continuous, the values of covariates could be averaged for groups of realized and un-realized edges and the difference taken as a test statistic. These ideas have some possibilities but would seem to be less solidly supported than the use of triangles connected with social science theory in Question 1.

Question 4. Within the context of model comparison, there are a number of approaches that could be considered. First, note that an Erdos-Renyi model is a member of the class of binary glms of Question 2. In a situation with continuous covariates and an intercept term included in the model, an Erdos-Renyi model results from all regression coefficients being 0 other than the intercept. In a block model formulated with an intercept and other "effects" parameters, the same is true. In a block model formulated without an intercept (or overall mean) term, an Erdos-Renyi model results from equality of all the parameters. This suggests the use of likelihood ratio tests or confidence intervals for parameters as a way to select between the reduced Erdos-Reyni model and the full model of Question 2. A potential drawback for this approach is that the observed graph (Figure 1) contains only 10 nodes, which results in 45 random variables for the glm model of Question 2. This may or may not give us pause in using asymptotic results for inference, most likely depending on the number of covariates included in the model.

Another possibility would be to use the fact that the binary response distributions form an exponential dispersion family to motivate the use of the difference in deviances for reduced Erdos-Renyi and full glm models. This results in a difference of (log) likelihoods for models fitted to the same data. To determine a reference distribution for this statistic, one could embed its use in a Monte Carlo procedure similar to that of Question 1. The difference with that algorithm would occur in calculation of the test statistic and determination of the $p-$value. Specifically, step (b) (ii) in that algorithm would be replaced by

(b) ii Estimate both Erdos-Renyi and glm models using the data $\boldsymbol{y}^*_m$, and use these model fits to produce estimates of the θ_i. Denote those estimates for the Erdos-Renyi model as $\boldsymbol{\theta}^*_{R,m}$ and those for the full glm as $\boldsymbol{\theta}^*_{F,m}$ and compute the statistic

$$T^*_m = \log[f(\boldsymbol{y} *_m |\boldsymbol{\theta}^*_{F,m}) - f(\boldsymbol{y}^*_m|\boldsymbol{\theta}^*_{R,m})].$$

Note that there is no need for scaling here as we are determining the reference distribution through simulation.

Since only large values of this statistic indicate preference for the full glm over the reduced Erdos-Renyi model, compuation of the Monte Carlo $p-$value in step (c) of the algorithm can also be modified to be

(c) Compute the test statistic for the observed data as

$$T = \log[f(\boldsymbol{y}|\boldsymbol{\theta}_F) - f(\boldsymbol{y}_m|\boldsymbol{\theta}_R)].$$

and the Monte Carlo $p-$value as

$$p = \sum_{m=1}^{M} I(T^*_m \leq T).$$

Question 5. The danger in viewing model assessment as only a matter of model selection is that none of the models being compared may be "adequate" in the sense of being reasonable generators of the observed data. That is, we may end up only selecting the best among a set of bad models. The danger in viewing model assessment as only a matter of the ability of the model to generate reasonable looking data is that it may ignore connections between model components and the intent of analysis. The classic but trivial case of this is that, ignoring an intent to relate the expectations of response variables to covariate values in a linear regression, a one-sample model with constant mean may well generate data sets that look for many purposes the same as the observed data. And, similarly to the situation of Question 3, it may not be straightforward to detect a relation between responses and covariates without some idea of what that relation might look like. A considerably more involved example would be four-nearest versus eight-nearest neighbor spatial models on a regular lattice, which we discussed in class but is not expected as a part of this answer.

Question 6. In a non-Bayesian analysis the most readily available approach for estimation is the use of a composite likelihood, the simplest of which in this case would be Besag's original psuedo-likelihood. Under this approach, parameter estimates would be found by maximizing the objective function

$$\ell_P(\beta_0, \beta_1, \eta) = \sum_{i=1}^{k} \log[f(y_i|p_i)],$$

where the p_i are defined as functions of β_0, β_1, and η in expressions (2) and (3) of the prelim question.

Question 7. There are any number of possibilities here. Two that I expect in answer to this question (only one in any given answer) are:

(a) Compute an interval for η through use of the asymptotics of composite likelihood (specifically for Besag's psuedo-likelihood), namely asymptotic normality. It would be reasonable here to appeal to the asymptotic context of a repeating lattice, given the problem definition with $n = 10$ fixed nodes in the observed graph. Thus, although one could compute a covariance matrix for this purpose through the use of a particular limit theorem (e.g., that of Mardia and Marshall) for Besag's psuedo-likelihood, or through application of a block bootstrap or resampling strategy, the most easily implemented possibility would be to use Godamde information.

(b) Compute an interval for η through the use of a parametric bootstrap procedure, and a percentile bootstrap interval in particular. Here, an important point to consider is how to simulate data sets from a fitted model. Use of the joint distribution as given in expression (4) of the exam question is not feasible, and the only other distributions specified in the model are the conditionals. Because these are *full conditional* distributions, however, the use of a Gibbs Sampling algorithm with estimated parameter values would be a way to simulate realizations from the model.

Question 8. Some of the answers that would be reasonable for this question are given here. Not all of these need be mentioned in an adequate answer, and other adequate possibilities may exist if they are clearly explained.

Either of the approaches suggested in Question 7 may present both some computational challenges and difficulties with justification. Relative to Question 7(a), computation of quantities that can play the role of a covariance matrix in a limiting distribution can become involved in terms of "book keeping" and ensuring that the appropriate terms are used in summations (although the use of Godambe information would make this less of a burden than the other options). A major concern with this approach using any technique to determine a covariance matrix is the small sample size of this problem ($k = 45$). Using the rough guide that our "usual" notions of sample size need to be squared for models that involve complex dependence structures, this is an extremely small sample to try to justify the use of asymptotic results for inference.

In terms of the answer to Question 7(b), implementation of parametric bootstraph involves not only the simulation of data sets, but also the automation of estimation procedures to use with Monte Carlo data sets. This could pose a considerable computational challenge. Profiling of the dependence parameter η might help, although it would increase the sheer number of computations to be conducted. If a percentile interval procedure were used, there would be the question of whether there exists a transformation of the parameter estimator that produces symmetry in (here the individual marginal) sampling distributions of the components of the estimator. Although this is almost never checked in particular applicaitons, it is worth remembering that it remains an issue.

Question 9. The major difficulty with a Bayesian approach to this problem is that the likelihood is not available in closed form, as indicated in the question immediately following expression (4).

Question 10. Although the scientific basis for formulating the LSG model has implications for how a realized graph might be structured, it does not lead to a simply formulated characteristic of graphs that could be used as a test statistic in the procedure of Question 1. Thus, it would seem difficult to approach this problem from the viewpoint of assessing the adequacy of an Erdos-Renyi model in isolation from any alternative. And, although the Erdos-Renyi model might be considered a reduced version of the LSG model (set $\eta = 0$ and $\beta_1 = 0$), likelihood results that can be directly applied are elusive. Use of a likeihood ratio-type (or score, or Wald) test would require the identification of specific asymptotic results that are applicable (results for composite likelihood have focused on asymptotic normality rather than behavior of likelihoods, but there are some likelihood ratio type results in Guyon which could be looked at). Even if an appropriate result could be located, there would be the same sample size concern identified in (the answer to) Question 4. At the same time, parametric bootstrap procedures are available for interval estimation of η under the LSG model (Question 7), as is Monte Carlo based comparison of an independence model using covariate information with an Erdos-Renyi model (Question 2). Since setting $\eta = 0$

in the LSG model leads to the independence covariate model, these two procedures might be combined by first demonstrating that the covariate model is preferred to the Erdos-Renyi model using the Monte Carlo procedure, and then demonstrating that 0 is not include in an interval for η under the LSG model with a parametric bootstrap.

Bike sharing systems are becoming a popular variant to traditional bicycle rentals. Part of a number of sustainability initiatives, Iowa State University has launched a bike share pilot project in 2013. This program is heavily administered by ISU students, and many of the initiatives have been developed as class projects. Students in Community and Regional Planning (CRP) have been asked to design a model to predict the daily volume of bicycle rentals from environmental and seasonal variables, using the data collected at campuses similar to ISU's. Even though the number of predictors is fairly large, the students have been asked to consider first just the effect of humidity on daily bike rental volume. You, as a Statistics graduate student, are asked to help the CRP students with their data analysis.

Assume the data are in the form of independent pairs, $(Y_j, x_j),\ j = 1, \ldots, n$, where $n = 350$, x_j is some normalization of the humidity variable (actual humidity divided by 100), and Y_j are daily volumes (treated as continuous responses here). The CRP students are very confident that

$$E(Y_j|x_j) = f(x_j, \boldsymbol{\beta});\quad \text{Var}(Y_j|x_j) = \sigma^2 x_j^{2\theta},\ j = 1, \ldots, n, \tag{1}$$

where $f(x)$ is a function that is decreasing in x, and $\theta \geq 0$ is an unknown scalar. Based on information from other studies, students are also sure that it is reasonable to assume that $\epsilon_j = \{Y_j - f(x_j, \boldsymbol{\beta_0})\}/\{\sigma_0 x_j^{\theta_0}\}$ are independent and identically distributed (i.i.d.), where $(\boldsymbol{\beta_0}, \sigma_0, \theta_0)$ are the true values of $(\boldsymbol{\beta}, \sigma, \theta)$. However, the only information they have about the shape of the distribution of ϵ_j is that it probably is symmetric.

You are asked to estimate $\boldsymbol{\beta}$ by generalized least squares. In order to form weights, you need to first estimate θ.

Part I

1. Explain, based on theoretical results, the relationship between the precisions with which you estimate $\boldsymbol{\beta}$ and θ.

2. One possible approach to estimating θ is to consider logarithms of absolute residuals as responses in estimating equations. Explain how this method would be used here, and how it boils down to a simple linear regression where θ is the slope.

3. Briefly explain the drawbacks of the method in Problem **2**. on the estimation of $\boldsymbol{\beta}$ when data are normally distributed.

4. Another possible approach to estimating θ is to consider the squared residuals as the responses. Explain how this method would be used here.

5. Explain why, for the model described in equation (1), the method in Problem **4**. might be preferred to the method in Problem **2**.

Part II

You have decided to estimate θ based on squared residuals. You need to test the null hypothesis $H_0 : \theta = 0$ against the alternative $H_a : \theta \neq 0$ to assess whether or not there is evidence, in the context of model (1), supporting the claim that variance is not constant.

6. Explain how to conduct this test of hypothesis if you assumed that $Y_j|x_j$ is normally distributed, using results of the standard large sample theory.

7. Explain why the development of the hypothesis test in Problem **6**. will not be valid if the normality condition is not satisfied.

8. The CRP students want to use the fitted model to calculate prediction intervals for values of the response that might be observed at different settings of x. They have found an online program that calculates approximate $100(1 - \alpha)\%$ prediction intervals for nonlinear models based on large sample theory under the assumption that the conditional variance of the response given x is constant for all j. Explain why this is or is not a reasonable procedure for use in the present context.

Part III

CRP students also have data from a much smaller study, with $n = 9$, conducted at ISU. They believe that model (1) and all of the assumptions about it should still hold for these data. They plan to estimate $(\boldsymbol{\beta}, \sigma, \theta)$ using generalized least squares for $\boldsymbol{\beta}$ and employ one of the methods in Problems **2**. or **4**. to estimate θ.

9. Explain why the application of standard large sample theory does not work in this situation. In particular, comment on which aspects of the estimation may be affected and how.

10. Propose a possible remedy that will allow you to construct reasonable confidence intervals for the components of $\boldsymbol{\beta}$.

Bike sharing systems are becoming a popular variant to traditional bicycle rentals. Part of a number of sustainability initiatives, Iowa State University has launched a bike share pilot project in 2013. This program is heavily administered by ISU students, and many of the initiatives have been developed as class projects. Students in Community and Regional Planning (CRP) have been asked to design a model to predict the daily volume of bicycle rentals from environmental and seasonal variables, using the data collected at campuses similar to ISU's. Even though the number of predictors is fairly large, the students have been asked to consider first just the effect of humidity on daily bike rental volume. You, as a Statistics graduate student, are asked to help the CRP students with their data analysis.

Assume the data are in the form of independent pairs, $(Y_j, x_j),\ j = 1, \ldots, n$, where $n = 350$, x_j is some normalization of the humidity variable (actual humidity divided by 100), and Y_j are daily volumes (treated as continuous responses here). The CRP students are very confident that

$$E(Y_j|x_j) = f(x_j, \boldsymbol{\beta});\ \ \mathrm{Var}(Y_j|x_j) = \sigma^2 x_j^{2\theta},\ j = 1, \ldots, n, \tag{1}$$

where $f(x)$ is a function that is decreasing in x, and $\theta \geq 0$ is an unknown scalar. Based on information from other studies, students are also sure that it is reasonable to assume that $\epsilon_j = \{Y_j - f(x_j, \boldsymbol{\beta_0})\}/\{\sigma_0 x_j^{\theta_0}\}$ are independent and identically distributed (i.i.d.), where $(\boldsymbol{\beta_0}, \sigma_0, \theta_0)$ are the true values of $(\boldsymbol{\beta}, \sigma, \theta)$. However, the only information they have about the shape of the distribution of ϵ_j is that it probably is symmetric.

You are asked to estimate $\boldsymbol{\beta}$ by generalized least squares. In order to form weights, you need to first estimate θ.

Part I

1. Explain, based on theoretical results, the relationship between the precisions with which you estimate $\boldsymbol{\beta}$ and θ.

   Answer: According to large sample theory, there is no effect of estimating θ or the the method you use to estimate it on the precision with which you would estimate $\boldsymbol{\beta}$. The sample size is fairly large ($n = 350$), therefore it is very likely that it makes very little difference on the precision of $\boldsymbol{\beta}$ which method you use to estimate θ. However, you can never be sure in any given problem what the implications of asymptotic theory are, so it may be possible that the method you use to estimate θ will matter in terms of how well you can estimate $\boldsymbol{\beta}$. For samples not large enough, the precision with which you can estimate $\boldsymbol{\beta}$ is directly related to that with which you estimate θ, so it is worth worrying about the procedure to estimate θ, to be safe.

2. One possible approach to estimating θ is to consider logarithms of absolute residuals as responses in estimating equations. Explain how this method would be used here, and how it boils down to a simple linear regression where θ is the slope.

   Answer: The method based on logarithms of absolute residuals regresses the log of the absolute value of $|r_j| = |Y_j - f(x_j, \hat{\boldsymbol{\beta}}^*)|$, where $\hat{\boldsymbol{\beta}}^*$ is a preliminary estimator for $\boldsymbol{\beta}$ (such as OLS or *a previous GLS estimator) against the logarithm of the model standard deviation, which, in your case is* $\log \sigma + \theta \log x_j$. *Thus, for your model, it boils down to a simple linear regression, where the estimator of* θ *is the estimator of slope.*

3. Briefly explain the drawbacks of the method in Problem **2**. on the estimation of $\boldsymbol{\beta}$ when data are normally distributed.

Answer: Since the variance model does not depend on $\boldsymbol{\beta}$, *symmetric distribution and* ϵ_j *are i.i.d., the method based on logarithms of absolute residuals can be very inefficient relative to the other methods when the data are in fact normally distributed or even when they are prone to "unusual" observations. So, if you use this method, be aware that its simplicity is offset by relative imprecision, which can translate into poor estimation of* $\boldsymbol{\beta}$.

4. Another possible approach to estimating θ is to consider the squared residuals as the responses. Explain how this method would be used here.

Answer: The method based on squared residuals is like a weighted regression with $|r_j|^2$ *as the "responses" and the variance* $\sigma^2 x_j^{2\theta}$ *as the "mean model."*

5. Explain why, for the model described in equation (1), the method in Problem **4**. might be preferred to the method in Problem **2**.

Answer: The method based on squared residuals can be derived by making the assumption that $Y_j|x_j$ *is normal and because, even if it is not, it will always produce a consistent estimator for* θ. *The method based on squared residuals is not hard to implement for a model like (1) (it can be carried out using standard nonlinear regression software), so it is not appreciably more complicated to implement than the method based on logarithms. Given its inferior precision relative to the method using the squared residuals, it is recommended to use the latter.*

Part II

You have decided to estimate θ based on squared residuals. You need to test the null hypothesis $H_0 : \theta = 0$ against the alternative $H_a : \theta \neq 0$ to assess whether or not there is evidence, in the context of model (1), supporting the claim that variance really is not constant.

6. Explain how to conduct this test of hypothesis if you assumed that $Y_j|x_j$ is normally distributed, using results of the standard large sample theory.

Answer: Given the assumption of normality, you can construct either a Wald test statistic or a likelihood-based score type statistic that should have approximately a chi-square distribution with 1 degree of freedom.

7. Explain why the development of the hypothesis test in Problem **6**. will not be valid if the normality condition is not satisfied.

Answer: If, in fact, your data are not normally distributed (but are symmetric, as you assumed, just with heavier tails), the test statistics mentioned above will no longer have a χ_1^2 *sampling distribution. Instead, the sampling distribution will be a scaled version of* chi_1^2, *with* *the scale factor depending on the excess kurtosis of the true distribution of $Y_j|x_j$. If you go ahead and carry out your test assuming the test test statistic does follow a χ^2_1 distribution, the results would be erroneous, as the p-values you would obtain will not be correct.*

8. The CRP students want to use the fitted model to calculate prediction intervals for values of the response that might be observed at different settings of x. They have found an online program that calculates approximate $100(1-\alpha)\%$ prediction intervals for nonlinear models based on large sample theory under the assumption that the conditional variance of the response given x is constant for all j. Explain why this is or is not a reasonable procedure for use in the present context.

Answer: This problem has to do with the implications of getting the variance model incorrect for the validity of prediction intervals. In a sample size as large as yours, the major source of variation that determines the width of the precision intervals is the variance of $Y_j|x_j$ itself–the variation due to fitting the model is dominated by this source. If the variance of $Y_j|x_j$ is not constant but instead has the form postulated in (1), pretending that it is constant and using this program will lead to intervals that are too wide in the regions of the rename of x where $\sigma^2 x^{2\theta}$ is small and too narrow in regions where $\sigma^2 x^{2\theta}$ is large. This is because the width of valid prediction intervals will not be constant if the variance is not constant.

In addition, regardless of whether the variance is constant or not, standard formulae for prediction intervals are based on the assumption that the distribution of $Y_j|x_j$ is approximately normal. If the distribution deviates from normality, all of these intervals will probably be incorrect.

Part III

CRP students also have data from a much smaller study, with $n = 9$, conducted at ISU. They believe that model (1) and all of the assumptions about it should still hold for these data. They plan to estimate $(\boldsymbol{\beta}, \sigma, \theta)$ using generalized least squares for $\boldsymbol{\beta}$ and employ one of the methods in Problems **2**. or **4**. to estimate θ.

9. Explain why the application of standard large sample theory does not work in this situation. In particular, comment on which aspects of the estimation may be affected and how.

Answer: With such a small sample size, it is very likely that the fact that you have estimated θ will probably affect the true precision with which you estimate $\boldsymbol{\beta}$. The standard large sample theory assumes that how you estimated θ has no effect, so the standard errors you calculate based on it will probably be too small relative to the true extent of sampling variation inherent in estimation of $\boldsymbol{\beta}$, because they do not take the additional variation due to estimating θ into account. This will lead to confidence intervals being too narrow to achieve the state level of coverage.

10. Propose a possible remedy that will allow you to construct reasonable confidence intervals for the components of $\boldsymbol{\beta}$.

Answer: A possible alternative is to instead use the bootstrap to obtain standard errors and form confidence intervals. It has been shown that the bootstrap standard errors correct for the effect of estimating θ that the standard large sample theory does not capture, so this may produce a more reliable confidence interval— all assuming that your variance model is correctly specified.