PhD Prelim Exam
METHODS

(Majors and Co-majors)

Summer 2012
(Given on 7/10/12)

Data were collected from each of the fifty states in the U.S. to examine variables that might be related to per capita expenditures on public education. In order to track changes over time, data were collected on the same variables in 1965, 1970 and 1975. The variables are:

- Year year in which the data were collected (1965, 1970, 1975)
- Y per capita expenditure on public education by the state (dollars per person)
- X_1 per capita income in the state (dollars per person)
- X_2 percentage of the state population under 18 years of age
- X_3 percentage of the state population living in urban areas

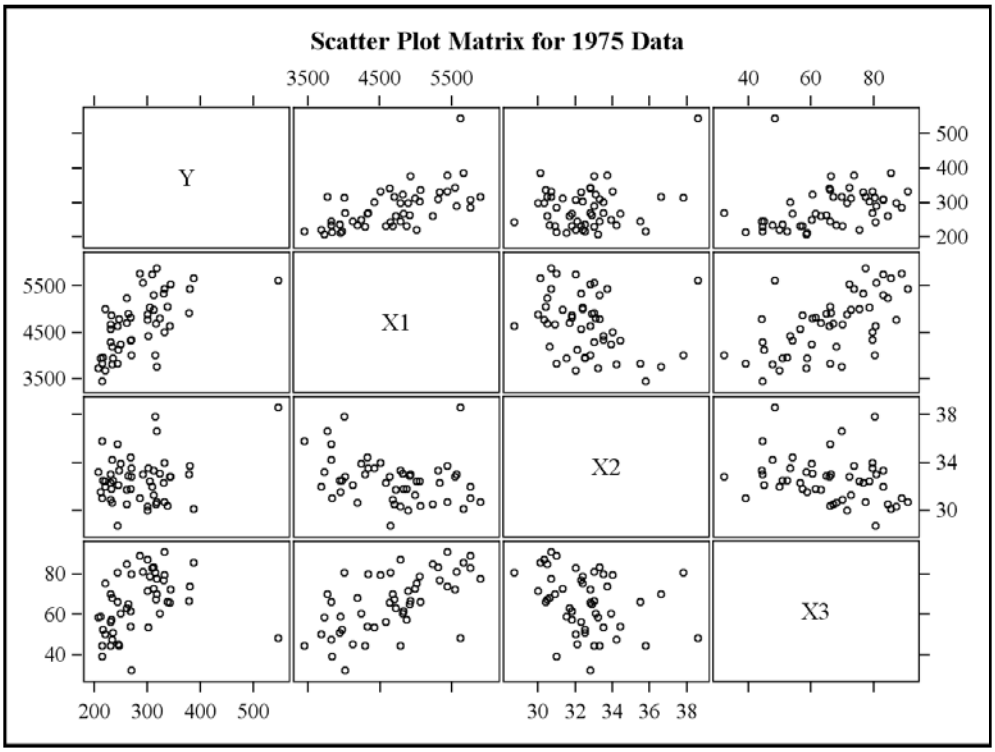
To help you to better understand the data, a listing of the 1975 data for the fifty states is displayed below.

Obs	State	Year	Y	X1	X2	X3
1	ME	1975	235	3944	32.5	50.8
2	NH	1975	231	4578	32.3	56.4
3	VT	1975	270	4011	32.8	32.2
4	MA	1975	261	5233	30.5	84.6
5	RI	1975	300	4780	30.3	87.1
6	CT	1975	317	5889	30.7	77.4
7	NY	1975	387	5663	30.1	85.6
8	NJ	1975	285	5759	31.0	88.9
9	PA	1975	300	4894	30.0	71.5
10	OH	1975	221	5012	32.4	75.3
11	IN	1975	264	4908	32.9	64.9
12	IL	1975	308	5753	32.0	83.0
13	MI	1975	379	5439	33.7	73.8
14	WI	1975	342	4634	32.8	65.9
15	MN	1975	378	4921	33.0	66.4
16	IA	1975	232	4869	31.8	57.2
17	MO	1975	231	4672	30.9	70.1
18	ND	1975	246	4782	33.3	44.3
19	SD	1975	230	4296	33.0	44.6
20	NB	1975	268	4827	31.8	61.5
21	KS	1975	337	5057	30.4	66.1
22	DE	1975	344	5540	32.8	72.2
23	MD	1975	330	5331	32.3	76.6
24	VA	1975	261	4715	31.7	63.1
25	WV	1975	214	3828	31.0	39.0

Obs	State	Year	Y	X1	X2	X3
26	NC	1975	245	4120	32.1	45.0
27	SC	1975	233	3817	34.2	47.6
28	GA	1975	250	4243	33.9	60.3
29	FL	1975	243	4647	28.7	80.5
30	KY	1975	216	3967	32.5	52.3
31	TN	1975	212	3946	31.5	58.8
32	AL	1975	208	3724	33.2	58.4
33	MS	1975	215	3448	35.8	44.5
34	AR	1975	221	3680	32.0	50.0
35	LA	1975	244	3825	35.5	66.1
36	OK	1975	234	4189	30.6	68.0
37	TX	1975	269	4336	33.5	79.7
38	MT	1975	302	4418	33.5	53.4
39	ID	1975	268	4323	34.4	54.1
40	WY	1975	323	4813	33.1	60.5
41	CO	1975	304	5046	32.4	78.5
42	NM	1975	317	3764	36.6	69.8
43	AZ	1975	332	4504	34.0	79.6
44	UT	1975	315	4005	37.8	80.4
45	NV	1975	291	5560	33.0	80.9
46	WA	1975	312	4989	31.3	72.6
47	OR	1975	316	4697	30.5	67.1
48	CA	1975	332	5438	30.7	90.9
49	AK	1975	546	5613	38.6	48.4
50	HI	1975	311	5309	33.3	83.1

This problem leads you through a sequence of analyses with a related set of eleven questions presented in three sets. Each set of questions is presented just after sufficient information has been provided to allow you answer them. Questions 1 through 4 are on the bottom of page 5. Questions 5 and 6 are on the top of page 8. Questions 7 through 11 are on page 13.

For each of the three years in which data were collected (1965, 1970, 1975), the researchers constructed a scatterplot matrix of the Y, X_1, X_2, X_3 variables, computed a correlation matrix, and computed least squares estimates of the regression coefficients for the regression of Y on X_1, X_2, and X_3. The results are shown below for the data from 1975.



Pearson Correlation Coefficients for the 1975 Data

The estimated correlation coefficient is the top number in each box.

The p-value for testing H0: $\rho = 0$ against HA: $\rho \neq 0$ is the second number in each box.

	Y	X1	X2	X3
Y	1.00000	0.60830 <.0001	0.26843 0.0595	0.32212 0.0225
X1	0.60830 <.0001	1.00000	-0.29704 0.0362	0.62194 <.0001
X2	0.26843 0.0595	-0.29704 0.0362	1.00000	-0.28652 0.0437
X3	0.32212 0.0225	0.62194 <.0001	-0.28652 0.0437	1.00000

Notation: Throughout this question the index i=1,2,...,50 is used to indicate the states as numbered in the table on page 1, and the index j=1,2,3 is used to indicate the year of the study (j=1 for 1965, j=2 for 1970, j=3 for 1975). The following subscript notation is used to indicate the state and year associated with values of the response variable and the explanatory variables:

Y_{ji} denotes per capita expenditure on public education (dollars per person) by the i-th state during the j-th year (j=1 for 1965, j=2 for 1970, j=3 for 1975)

X_{1ji} denotes per capita income (dollars per person) within the i-th state during the j-th year

X_{2ji} denotes the percentage of the population under 18 years of age in the i-th state during the j-th year

X_{3ji} percentage of the population living in rural areas in the i-th state during the j-th year

Separate regression models were fit to the data from the fifty states for each of the three years of the study. We will first consider the results for the 1975 data, the third year (j=3) of the study. The regression model is

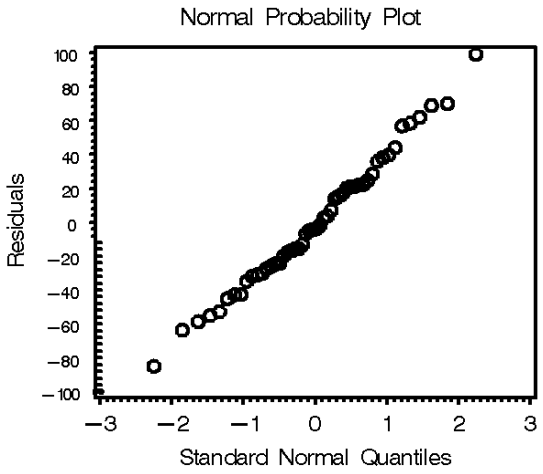
$$Y_{3i} = \beta_{03} + \beta_{13}X_{13i} + \beta_{23}X_{23i} + \beta_{33}X_{33i} + \varepsilon_{3i} \qquad \text{(Model 3)}$$

where β_{03}, β_{13}, β_{23}, and β_{33} are unknown regression coefficients and ε_{3i} denotes a random error. Assume that the random errors are independent and normally distributed with mean zero and variance σ_3^2. The least squares estimates of the regression coefficients and the ANOVA table are shown below. Plots of residuals and partial residual plots are also displayed.

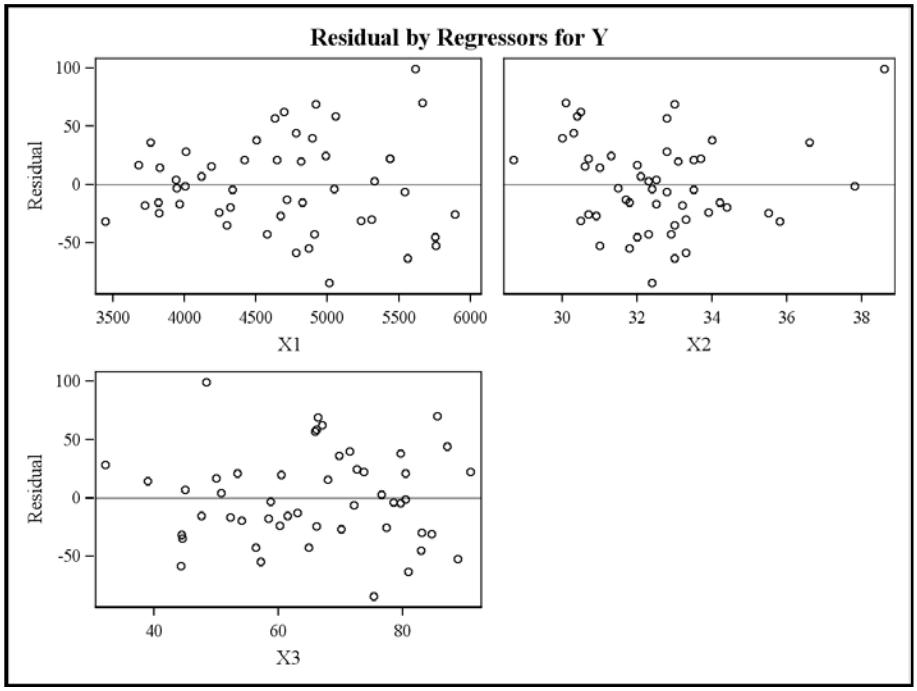
Parameter Estimates for the 1975 Data

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-556.56804	123.19525	-4.52	<.0001	0
X1	1	0.07239	0.01160	6.24	<.0001	1.67278
X2	1	15.52054	3.14672	4.93	<.0001	1.11747
X3	1	-0.04269	0.51393	-0.08	0.9342	1.66159

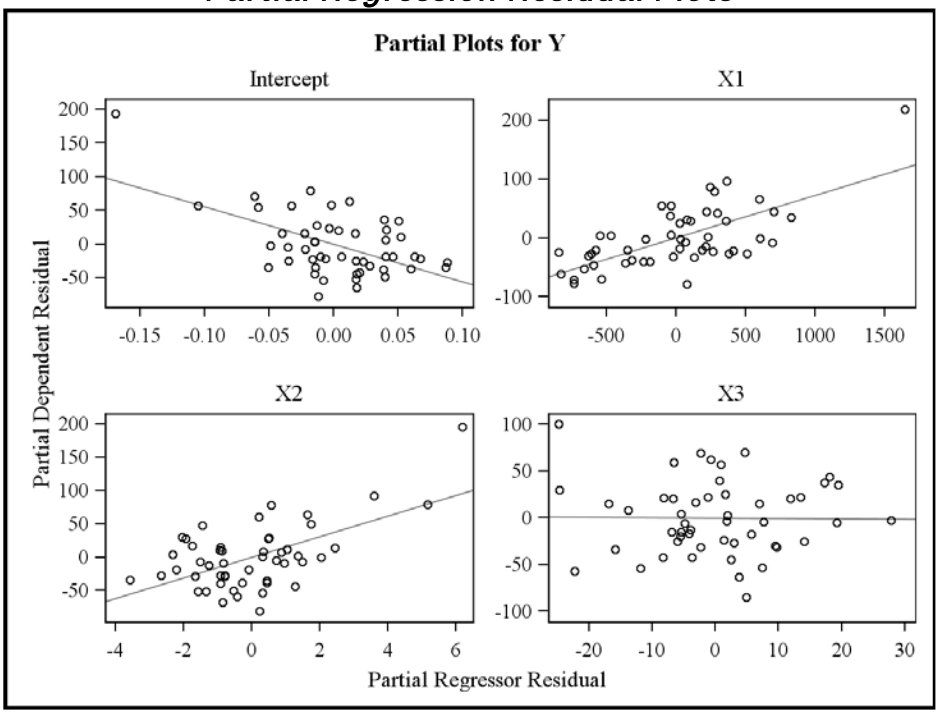
Analysis of Variance for the 1975 Data					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	109020	36340.000	22.19	<.0001
Error	46	75348	1637.991		
Corrected Total	49	184368			



Residual Plots



Partial Regression Residual Plots



1. For the 1975 data, what do the scatterplot matrix and the correlation results tell you about possible associations among the variables: per capita expenditure on public education (Y), per capita income (X_1), percentage of the population under 18 (X_2), and percentage of the population in urban areas (X_3) for that year?

2. The least squares estimate of β_{13} is reported as 0.07239 with a standard error of 0.0116, and the corresponding p-value is less than 0.0001. Carefully interpret this regression coefficient and the results of the test of significance in the context of the study of per capita expenditures on public education. As part of your answer, indicate whether you would conclude that increasing per capita income in a state would cause an increase in per capita expenditures on public education and provide relevant justification for your conclusion.

3. State the conclusions you would reach from inspection of the normal probability plot of the residuals and the plots of the residuals against the values of the individual explanatory variables that are displayed on page 4.

4. State the conclusions you would reach from inspection of the partial residual plots displayed at the top of this page. As part of your answer, discuss the point that appears in the upper right corner of the partial residual plot for per capita income (X_1). In particular, would you expect that observation to have high leverage or to be highly influential?

The regression model for the 1970 data, the second (j=2) year of the study, is

$$Y_{2i} = \beta_{02} + \beta_{12}X_{12i} + \beta_{22}X_{22i} + \beta_{32}X_{32i} + \varepsilon_{2i} \qquad \text{(Model 2)}$$

where the ε_{2i} are independent and normally distributed random errors with mean zero and variance σ_2^2. The results of fitting this model to the 1970 data are shown below.

Parameter Estimates for the 1970 Data					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-281.97802	68.83577	-4.10	0.0002
X1	1	0.07661	0.00959	7.99	<.0001
X2	1	8.21178	1.68404	4.88	<.0001
X3	1	-0.96012	0.36158	-2.66	0.0108

Analysis of Variance for the 1970 data					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	69107	23036	29.21	<.0001
Error	46	36276	788.602		
Corrected Total	49	105383			

The model for the 1965 data, the first (j=1) year of the study, is

$$Y_{1i} = \beta_{0,65} + \beta_{11}X_{11i} + \beta_{21}X_{21i} + \beta_{31}X_{31i} + \varepsilon_{1i} \qquad \text{(Model 1)}$$

where the ε_{1i} are independent and normally distributed random errors with mean zero and variance σ_1^2. The results of fitting this model to the 1965 data are shown below.

Parameter Estimates for the 1965 data					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	1	-11.40463	28.97616	-0.39	0.6957
X1	1	0.04493	0.00767	5.86	<.0001
X2	1	0.66223	0.48834	1.36	0.1817
X3	1	-0.28954	0.19293	-1.50	0.1403

Analysis of Variance for the 1965 data					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	10134	3377.871	13.71	<.0001
Error	46	11332	246.354		
Corrected Total	49	21466			

The researchers also fit the following regression model to the combined data set for all three years of the study:

$$Y_{ji} = \beta_{0,all} + \beta_{1,all}X_{1ji} + \beta_{2,all}X_{2ji} + \beta_{3,all}X_{3ji} + \varepsilon_{ji} \qquad \text{(Model 4)}$$

where the ε_{ji} are independent and normally distributed random errors with mean zero and variance σ^2. Note that this model assumes that values of regression coefficients do not change across the three years of the study and the variances of the random errors are homogeneous across years.

The results of fitting Model 4 to the entire set of data are displayed below.

Parameter Estimates for Model 4						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	-139.39836	40.87997	-3.41	0.0008	0
X1	1	0.07995	0.00347	23.04	<.0001	2.18279
X2	1	2.93902	0.84938	3.46	0.0007	1.94653
X3	1	-0.63938	0.22243	-2.87	0.0047	1.30465

Analysis of Variance for Model 4					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	1121821	373940	289.36	<.0001
Error	146	188675	1292.297		
Corrected Total	149	1310497			

5. Assume Models 1, 2, and 3 provide appropriate descriptions of the data for 1965, 1970, and 1975, respectively. Also assume that all 150 random errors are mutually independent and normally distributed with mean zero and constant variance σ^2. Use the information provided on the first seven pages of this question to perform a test of the null hypothesis that the true regression model is the same for all three years, i.e., test the null hypothesis

$$H_0: \beta_{01} = \beta_{02} = \beta_{03} \text{ and } \beta_{11} = \beta_{12} = \beta_{13}$$
$$\text{and } \beta_{21} = \beta_{22} = \beta_{23} \text{ and } \beta_{31} = \beta_{32} = \beta_{33}$$

against the general alternative that the null hypothesis is not completely true. Show your work and state your conclusion.

6. One assumption you were instructed to make in performing the test in question 5 is that the error variances are homogeneous across years. Assuming normal distributions and that the 150 random errors are independent, use the information provided on the first seven pages of this problem to assess the assumption of homogeneous error variances across years. Show your work and state your conclusions.

The researchers considered other models that allow the regression coefficients and the error variances to change across time and also allow for nonhomogeneous variances and correlation among observations taken from the same state in different years. Exactly as in Models 1, 2, and 3 considered before separately, conditional means for per capita expenditure on public education are

$$\begin{aligned} E(Y_{1i}) &= \beta_{01} + \beta_{11}X_{11i} + \beta_{21}X_{21i} + \beta_{31}X_{31i} \quad \text{for data from 1965} \\ E(Y_{2i}) &= \beta_{02} + \beta_{12}X_{12i} + \beta_{22}X_{22i} + \beta_{32}X_{32i} \quad \text{for data from 1970} \\ E(Y_{3i}) &= \beta_{03} + \beta_{13}X_{13i} + \beta_{23}X_{23i} + \beta_{33}X_{33i} \quad \text{for data from 1975} \end{aligned} \qquad \text{(Model 5)}$$

The researchers considered seven models for the covariance matrix for the random errors, $\varepsilon_{ji} = Y_{ji} - E(Y_{ji})$. Let

$$\underset{\sim}{\varepsilon}_i = \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \\ \varepsilon_{3i} \end{pmatrix} \quad i=1,2,3,....,50$$

denote the vector of random errors associated with the i-th state in 1965, 1970 and 1975, respectively. In the following it is assumed that $\underset{\sim}{\varepsilon}_1, \underset{\sim}{\varepsilon}_2, ..., \underset{\sim}{\varepsilon}_{50}$ are independent and identically distributed random vectors with a zero mean vector and covariance matrix Σ corresponding to one of seven possibilities. Different specifications of Σ yield different versions of Model 5 which are labeled as Model 5.1 through Model 5.7 in the output displayed below. The MIXED procedure in SAS, with the REML estimation option for parameters in the error covariance matrix, was used to fit each model. The REML log-likelihood is based on the

additional assumption that the error vectors are a simple random sample from a tri-variate normal distribution. Parameter estimates of the regression coefficients are given for each model along with REML estimates of parameters in the corresponding covariance matrix. The value of the REML log-likelihood and corresponding AIC and BIC values are also given for each model.

Model (5.1) $\Sigma_{5.1} = \begin{pmatrix} \sigma^2 & 0 & 0 \\ 0 & \sigma^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix}$ REML Estimate: $\hat{\sigma}^2 = 890.98$

REML log-likelihood = -716.96 AIC=1435.9 BIC=1438.8

Parameter Estimates for Model 5.1

Effect	DF	Estimate	Std. Error	t Value	Pr > |t|
Intercept1	1	-556.56804	90.86002	-6.13	<.0001
x11	1	0.04493	0.01458	3.08	0.0025
x12	1	0.66223	0.92871	0.71	0.4770
x13	1	-0.28954	0.36691	-0.79	0.4314
Intercept2	1	545.16342	106.26459	5.13	<.0001
x21	1	0.07661	0.01020	7.51	<.0001
x22	1	8.21178	1.79002	4.59	<.0001
x23	1	-0.96012	0.38434	-2.50	0.0137
Intercept3	1	274.59003	116.65789	2.35	0.0200
x31	1	0.07239	0.00856	8.46	<.0001
x32	1	15.52054	2.32079	6.69	<.0001
x33	1	-0.04269	0.37904	-0.11	0.9105

Model (5.2) $\Sigma_{5.2} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$ REML Estimates: $\hat{\sigma}^2 = 892.03$ $\hat{\rho}=0.4491$

REML log-likelihood = -704.60 AIC=1413.2 BIC=1417.0

Regression Parameter Estimates for Model 5.2					
Effect	**Estimate**	**Std. Error**	**DF**	**t Value**	**Pr > |t|**
Intercept1	-505.44	82.5331	137	-6.12	<.0001
x11	0.04182	0.01333	138	3.14	0.0021
x12	0.9466	0.8114	127	1.17	0.2455
x13	-0.1740	0.3398	137	-0.51	0.6094
Intercept2	481.99	87.4371	115	5.51	<.0001
x21	0.07193	0.009554	136	7.53	<.0001
x22	6.8439	1.6187	136	4.23	<.0001
x23	-0.9005	0.3636	133	-2.48	0.0145
Intercept3	283.67	87.6361	93.1	3.24	0.0017
x31	0.07047	0.008000	137	8.81	<.0001
x32	14.2709	2.1009	136	6.79	<.0001
x33	-0.06522	0.3598	132	-0.18	0.8564

Model (5.3) $\Sigma_{5.3} = \begin{pmatrix} \sigma^2 & \rho\sigma^2 & \rho^2\sigma^2 \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \\ \rho^2\sigma^2 & \rho\sigma^2 & \sigma^2 \end{pmatrix}$ REML estimates: $\hat{\sigma}^2 = 908.17$ $\hat{\rho}=0.5818$

REML log-likelihood: -699.60 AIC=1403.2 BIC=1407.0

Regression Parameter Estimates for Model 5.3					
Effect	**Estimate**	**Std. Error**	**DF**	**t Value**	**Pr > |t|**
Intercept1	-513.37	85.7618	134	-5.99	<.0001
x11	0.03797	0.01311	137	2.90	0.0044
x12	0.9093	0.7865	130	1.16	0.2498
x13	-0.08157	0.3343	136	-0.24	0.8076
Intercept2	493.59	92.8583	136	5.32	<.0001
x21	0.06857	0.009048	138	7.58	<.0001
x22	6.3588	1.4909	125	4.27	<.0001
x23	-0.8294	0.3485	136	-2.38	0.0187
Intercept3	315.11	77.1137	95.2	4.09	<.0001
x31	0.06933	0.008118	130	8.54	<.0001
x32	14.5713	2.1860	135	6.67	<.0001
x33	-0.01215	0.3672	121	-0.03	0.9737

Model (5.4) $\Sigma_{5.4} = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}$

REML Estimates: $\hat{\sigma}_1^2 = 246.35 \quad \hat{\sigma}_2^2 = 788.60 \quad \hat{\sigma}_3^2 = 1637.99$

REML log-likelihood: -698.55 AIC=1403.1 BIC=1408.9

Regression Parameter Estimates for Model 5.4					
Effect	**Estimate**	**Std. Error**	**DF**	**t -Value**	**Pr > |t|**
Intercept1	-556.57	123.20	46	-4.52	<.0001
x11	0.04493	0.007667	46	5.86	<.0001
x12	0.6622	0.4883	46	1.36	0.1817
x13	-0.2895	0.1929	46	-1.50	0.1403
Intercept2	545.16	126.56	51.1	4.31	<.0001
x21	0.07661	0.009593	46	7.99	<.0001
x22	8.2118	1.6840	46	4.88	<.0001
x23	-0.9601	0.3616	46	-2.66	0.0108
Intercept3	274.59	141.12	72.2	1.95	0.0556
x31	0.07239	0.01160	46	6.24	<.0001
x32	15.5205	3.1467	46	4.93	<.0001
x33	-0.04269	0.5139	46	-0.08	0.9342

Model (5.5) $\Sigma_{5.5} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$

REML Estimates: $\hat{\sigma}_1^2 = 254.32 \quad \hat{\sigma}_2^2 = 781.65 \quad \hat{\sigma}_3^2 = 1711.46 \quad \hat{\rho}$=0.5483

REML log-likelihood = -680.95 AIC=1369.9 BIC=1377.6

Regression Parameter Estimates for Model 5.5					
Effect	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept1	-386.45	107.92	56.8	-3.58	0.0007
x11	0.04256	0.006778	55.5	6.28	<.0001
x12	0.6776	0.4033	51.8	1.68	0.0990
x13	-0.2366	0.1738	55.6	-1.36	0.1788
Intercept2	375.99	104.89	55.4	3.58	0.0007
x21	0.07374	0.008583	58.4	8.59	<.0001
x22	5.5224	1.4265	57.7	3.87	0.0003
x23	-1.0106	0.3285	57.7	-3.08	0.0032
Intercept3	213.56	101.55	58.6	2.10	0.0398
x31	0.06937	0.01062	56.8	6.53	<.0001
x32	11.0961	2.7407	56.6	4.05	0.0002
x33	-0.2238	0.4817	55.6	-0.46	0.6441

Model (5.6) $\Sigma_{5.6} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 & \rho\sigma_2\sigma_3 \\ \rho^2\sigma_1\sigma_3 & \rho\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$

REML Estimates: $\hat{\sigma}_1^2 = 250.52$ $\hat{\sigma}_2^2 = 840.95$ $\hat{\sigma}_3^2 = 1618.16$ $\hat{\rho}$=0.6235

REML log-likelihood: -677.95 AIC=1363.9 BIC=1371.5

Regression Parameter Estimates for Model 5.6

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept1	-443.86	112.14	57.6	-3.96	0.0002
x11	0.04132	0.006728	55.1	6.14	<.0001
x12	0.6900	0.3991	51.2	1.73	0.0898
x13	-0.1906	0.1719	55.3	-1.11	0.2723
Intercept2	432.67	111.35	59.3	3.89	0.0003
x21	0.07207	0.008482	64	8.50	<.0001
x22	4.9737	1.3823	65.7	3.60	0.0006
x23	-0.9756	0.3281	61.3	-2.97	0.0042
Intercept3	293.72	95.7895	51.2	3.07	0.0035
x31	0.06973	0.01065	57.3	6.55	<.0001
x32	12.6215	2.8563	57.7	4.42	<.0001
x33	-0.1320	0.4838	56	-0.27	0.7859

Model (5.7) $\Sigma_{5.7} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{pmatrix}$

REML Estimates: $\hat{\sigma}_1^2 = 245.19$ $\hat{\sigma}_2^2 = 837.48$ $\hat{\sigma}_3^2 = 1657.23$ $\hat{\rho}_{12}$=0.5886 $\hat{\rho}_{13}$=0.3862 $\hat{\rho}_{23}$=0.6556

REML log-likelihood: -677.8 AIC=1367.6 BIC=1379.0

Regression Parameter Estimates for Model 5.7

Effect	Estimate	Std. Error	DF	t Value	Pr > |t|
Intercept1	-440.68	113.12	52.2	-3.90	0.0003
x11	0.04136	0.006777	53.1	6.10	<.0001
x12	0.7014	0.4063	45.6	1.73	0.0911
x13	-0.1933	0.1730	53.5	-1.12	0.2688
Intercept2	429.12	112.29	50.1	3.82	0.0004
x21	0.07122	0.008545	61.6	8.33	<.0001
x22	5.1232	1.4137	53	3.62	0.0007
x23	-0.9526	0.3321	56.3	-2.87	0.0058
Intercept3	286.39	93.7677	45	3.05	0.0038
x31	0.06900	0.01071	56.1	6.44	<.0001
x32	12.5988	2.8813	51.9	4.37	<.0001
x33	-0.1173	0.4891	52.5	-0.24	0.8114

7. The estimates of the regression parameters provided by the MIXED procedure in SAS are not the same for Models 5.1 through 5.7. Explain why the estimates of the regression coefficients differ for those models. In your answer, use $\underset{\sim}{Y}$ to indicate the 150x1 vector of observed per capita expenditures for the 50 state during 1965, 1970 and 1975, use Σ to indicate the true covariance matrix for $\underset{\sim}{Y}$, and use X to denote the corresponding 150x12 model matrix corresponding to the situation in which the elements of $\underset{\sim}{\beta}$ are organized as

$$\underset{\sim}{\beta} = \left(\beta_{01}\ \beta_{11}\ \beta_{21}\ \beta_{31}\ \beta_{02}\ \beta_{12}\ \beta_{22}\ \beta_{32}\ \beta_{03}\ \beta_{13}\ \beta_{23}\ \beta_{33}\right)'.$$

   Note that X is the same for all seven models, and the ordering of the rows in X and Y is not important. Is the estimation procedure used by the MIXED procedure in SAS equivalent to least squares estimation for any of these models?

8. Give formulas for the AIC and BIC statistics and explain how these quantities may be used to help identify one or more of the models that provide a good description of the data. Explain why the AIC ordering of Models 5.5 and 5.7 is not the same as the BIC ordering of those two models.

9. Perform a test to determine if Model 5.7 provides a significant improvement over Model 5.5 with respect to describing the covariance matrix for repeated measures of per capita expenditures on public education within states. State your conclusion.

10. The researchers are interested in estimating $\beta_{23} - \beta_{21}$ the difference in the regression coefficients for the percentage of the population under 18 years of age in 1975 and 1965. Using Model 5.7, show how to construct a 95% confidence interval for $\beta_{23} - \beta_{21}$. Use the notation established for question 7 to report any formulas you wish to present. Let $\underset{\sim}{Y}$ denote the vector of responses, let $\hat{\Sigma}$ denote the REML estimate of the covariance matrix for $\underset{\sim}{Y}$, and let X denote the model matrix corresponding to the situation in which the elements of $\underset{\sim}{\beta}$ are organized as $\underset{\sim}{\beta} = \left(\beta_{01}\ \beta_{11}\ \beta_{21}\ \beta_{31}\ \beta_{02}\ \beta_{12}\ \beta_{22}\ \beta_{32}\ \beta_{03}\ \beta_{13}\ \beta_{23}\ \beta_{33}\right)'$. Do not try to numerically evaluate the endpoints of the confidence interval, just explain how it should be constructed. If you use a procedure that relies on large sample distributional properties of estimators, do you expect that theory to be relevant in this application? Explain.

11. Up to this point you were allowed to assume that vectors of observations taken from different states are independent. The researchers suspect that this is not a valid assumption. Residents (and politicians) in neighboring states, for example, may have attitudes toward financing public education that are more similar than those for states that are farther apart. This would presumably induce a pattern of spatial correlations, with stronger correlations between per capita expenditures on public education for states that are closer together. Given that states have different geographical shapes and different areas, suggest a model that would account for spatial correlation among states in analyzing the data for the public education expenditure study.

1. Some points to consider:
   - Both the scatterplot matrix and the correlation results indicate moderate positive associations between expenditures on public education and the each of the three explanatory variables.
   - Per capita expenditure on public education appears to have the strongest association with per capita income.
   - There is a positive association between per capita income and percentage of the population who live in urban areas. There are negative associations between proportion of the population under 18 years of age and per capita income and percentage of the population living in urban areas.
   - The scatterplot matrix does not reveal any obvious curved relationships that would not be reflected by correlation coefficients.
   - The scatterplot matrix reveals one state, Alaska that has a much higher per capita expenditure on public education in 1975 than any of the other 49 states. Alaska is the state with the highest proportion of residents under 18 and it also has relatively high per capita income and a relatively low proportion of the population in urban areas.

2. The estimate of β_{13} is more than six standard errors above zero (and the p-value is less than .0001), indicating a significant positive association between expenditure on public education and per capita income, even after adjusting for correlations that expenditure on public education has with the percentage of the population in urban areas and the percentage of the population under age 18. The estimated value of 0.07239 suggests that after conditioning on the percentage of the population in urban areas and the percentage of population under 18 years of age, per capita expenditure on public education are expected to increase about 72 dollars for each $1000 increase in per capita income. Because the data are from an observational study, it would not be appropriate to make a causal inference. There are many possible uncontrolled factors, associated with both per capita income and per capita expenditure of public education, that could contribute to decisions on expenditures for public education that were not monitored in this study.

3. Some points to consider:
   - The points on the normal probability plot nearly lie along a straight line, indicating that the distribution of the random errors is approximately normal.
   - The residual plots exhibit no obvious trends that would indicate a curved relationship that is not explained by the proposed regression model.
   - Alaska exhibits a relatively large positive residual, but it is not large enough to completely discredit the proposed regression model.
   - There is some indication that variation about the regression model may be greater for states with larger per capita income.

4. Some points to consider:
   - The partial residual plots involving the intercept, per capita income, and proportion of the population under 18 all exhibit straight line patterns indicating that those variables are needed in the model. There are no curved patterns which would indicate that the proposed regression model fails to provide an adequate description of the relationship between per capita expenditure on public education and per capita income and percentage of population under 18.
   - The partial residual plot for the proportion of the population in rural areas exhibits random scatter about a horizontal line, indicating that changes in the percentage of the population in rural areas provides little information for predicting per capita expenditure on public education beyond the information provided by changes in per capita income and percentage of the population under 18 years of age.
   - The point that appears in the upper right corner of the partial residual plot for per capita income (X_1) indicates that there is one state with high per capita income relative to its percentage of residents in urban areas and its percentage of residents less than 18 years of age. This state could have moderately high leverage, but it is not a highly influential case with respect to the estimate of β_{13} because it coincides with the straight line trend exhibited by the observations for the other 49 states in that plot.

5. Comparing $$F = \frac{[SS_{error,model4} - (SS_{error,model1} + SS_{error,model2} + SS_{error,model3})]/8}{(SS_{error,model1} + SS_{error,model2} + SS_{error,model3})/138} = 9.22$$

   to the percentiles of a central F distribution with (8,138) degrees of freedom, yields a p-value less than 0.001. The data are inconsistent with the null hypothesis. The sets of regression coefficients do not appear to be the same for all three years. The coefficient for per capita income appears to have increased between 1965 and 1970. The coefficient for percent of the population younger than 18 appears to have increased every five years between 1965 and 1975.

6. In computing the denominator of the F-statistics in the solution to question 5, one may notice that the error mean square increased every five years between 1965 and 1975. Assuming that the observations are mutually independent and normally distributed, the error sums of squares for models 1, 2, and 3 are multiples of independent chi-square random variables. Consequently, ratios of the error mean squares for any two of the models will have a central F-distribution with (46,46) degrees of freedom when the error variances for the two models are the same. Compare

   $$F = \frac{MS_{error,model2}}{MS_{error,model1}} = 3.20 \qquad F = \frac{MS_{error,model3}}{MS_{error,model1}} = 6.65 \qquad F = \frac{MS_{error,model3}}{MS_{error,model2}} = 2.08$$

   to percentiles of the central F-distribution with (46,46) degrees of freedom, bearing in mind that each ratio has the largest mean square in the numerator. A Bonferroni adjustment could be considered for simultaneously performing three tests. It appears that variation about the regression lines increased over time as levels of per capita expenditures on public education increased.

7. This is a consequence of the estimation procedure. Initially model 5.1 is assumed and ordinary least squares estimates for model 5.1 are computed as $\hat{\underset{\sim}{\beta}}_{OLS} = (X'X)^{-1}X'\underset{\sim}{Y}$, where $\underset{\sim}{Y}$ is a 150x1 vector containing the observed values of per capita expenditures on public education in the states for 1965, 1970 and 1975 and X is the corresponding 150x12 model matrix. Then the likelihood function for 138 linearly independent combinations of the residuals from the ordinary least squares fit of model 5.1 to the data is maximized to obtain the REML estimates of the parameters in the error covariance matrix. Subsequently the estimated covariance matrix is used to obtain an approximation to a generalized least squares estimator for the regression coefficients $\hat{\underset{\sim}{\beta}}_{GLS} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}X'\hat{\Sigma}^{-1}\underset{\sim}{Y}$. Because the covariance matrix and its REML estimate is not the same for models 5.1 through 5.7, the value of $\hat{\underset{\sim}{\beta}}_{GLS} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}X'\hat{\Sigma}^{-1}\underset{\sim}{Y}$ is not the same for all of the models.

   Regression coefficient estimates for Model 5.1 are ordinary least squares estimates for the combined data. Regression coefficient estimates for Model 5.4 are ordinary least squares estimates for separately fitting Models 1, 2, and 3, respectively. Fitting the models separately allowed the error variances to change from year to year.

8. AIC and BIC are penalized information measures. For REML estimation AIC is evaluated as

   -2log(REML likelihood)+2(number of covariance parameters)

   and BIC is evaluated as

   -2log(REML likelihood)
   +(number of covariance parameters)*log(number of units).

   Generally, smaller values are assumed to reflect models that describe the data better without over fitting. For this study, the responding units are the 50 states and the penalty imposed by BIC is (number of covariance parameters)*log(50), which is 3.912(number of covariance parameters). This is almost double the penalty imposed by AIC. For these data, all of the models with heterogeneous error variances across years have relatively small AIC and BIC values and would be improvements over any of the covariance models with homogeneous error variances. Model 5.6 which has an AR(1) correlation pattern with heterogeneous error variances appears to provide the best fit. There appear to be positive correlations among observations taken on the same state and those correlations appear to be weaker for time points that are farther apart.

   The increase in the REML log-likelihood achieved by generalizing Model 5.5 to obtain Model 5.7 with unequal correlations to is -677.80-(-680.95)=3.15. This generalization requires three correlations parameters instead of one common

correlation parameter. Consequently, the change in the AIC values from Model 5.5 to Model 5.7 is a decrease of (2)(3.15)-(2)(2)=2.3, but the change in the BIC values is (2)(3.15)-(3.912)(2)=-1.524. Because BIC imposed a greater penalty for including more parameters in the model for the covariance matrix, the BIC increased while the AIC value decreased.

9. Likelihood ratio tests based on the REML log-likelihood values could be used to compare models. To determine if Model 5.7 is a significant improvement relative to model 5.5, for example, compare -2(680.95-677.80)=7.30 to percentiles of a central chi-square distribution with 2 degrees of freedom. This value is close to the .975 percentile, giving some evidence that Model 5.7 is an improvement over Model 5.5.

10. Let $\hat{\Sigma}$ denote the REML estimate of the covariance matrix for the combined set of 150 observations, using the model selected in question 8. A large sample estimate of the covariance matrix for $\hat{\underset{\sim}{\beta}}_{GLS} = \left(X'\hat{\Sigma}^{-1}X\right)^{-1}X'\hat{\Sigma}^{-1}\underset{\sim}{Y}$ is $\left(X'\hat{\Sigma}^{-1}X\right)^{-1}$. Assuming the elements of $\underset{\sim}{\beta}$ are organized as $\underset{\sim}{\beta} = (\beta_{01}\ \beta_{11}\ \beta_{21}\ \beta_{31}\ \beta_{02}\ \beta_{12}\ \beta_{22}\ \beta_{32}\ \beta_{03}\ \beta_{13}\ \beta_{23}\ \beta_{33})'$, a large sample standard error for $\hat{\beta}_{23} - \hat{\beta}_{21} = \underset{\sim}{C}'\hat{\underset{\sim}{\beta}}_{GLS} = (0\ 0\ -1\ 0\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0)'\hat{\underset{\sim}{\beta}}_{GLS}$ is $S_{\underset{\sim}{C}'\hat{\underset{\sim}{\beta}}_{GLS}} = \sqrt{\underset{\sim}{C}'\left(X'\hat{\Sigma}^{-1}X\right)^{-1}\underset{\sim}{C}}$. Consequently, an approximate 95%confidence interval for $\beta_{23} - \beta_{21} = \underset{\sim}{C}'\beta$ is $\underset{\sim}{C}'\hat{\underset{\sim}{\beta}}_{GLS} \pm t_{df,.975}S_{\underset{\sim}{C}'\hat{\underset{\sim}{\beta}}_{GLS}}$, where degrees of freedom can be obtained from the Kenward-Rogers method. This type of confidence interval tends to be too short and have less than the nominal coverage probability for small samples.

11. Students should provide reasonable comments on modeling potential spatial correlation. For example, one might consider computing the distance, d_{ik}, between the geographic centers of state i and state k and defining a correlation between those two states that is proportional to $\exp(-d_{ik})$ or $\exp(-d_{ik}^2)$ or some other reasonable function of distance. Such spatial correlation patterns can be accommodated by the MIXED procedure in SAS. There are often several ways to describe the same features of a data set. One might, for example, consider including dummy variables in the regression model to account for regional differences. Effects for a few regions such as, northeastern states, mid-Atlantic states, southeastern states, north-central states, southwestern states, great plains and western mountain states, and pacific coast states may account for most or all of potential spatial correlation patterns.

Part I

We consider a study of the toxicity of the chemical benzopyrene to juvenile and adult fish. Live fish were kept in tanks during the study. Each tank contained either 10 juvenile or 10 adult fish at the start of the study. Tanks were randomly assigned to a concentration of benzopyrene. Because juvenile fish were anticipated to be more susceptible to the toxicant, the concentrations of toxicant used for juveniles differed from those used for adults. After 60 days, the surviving fish were collected. The liver weight of each surviving fish was recorded. The numbers of tanks used for each combination of concentration and fish age is shown in Table 1.

			Concentration, μg/l (j)						
			0	2	5	10	20	50	100
			(1)	(2)	(3)	(4)	(5)	(6)	(7)
Fish Age (i)	Adult	(1)	2	0	0	2	2	2	2
	Juvenile	(2)	2	2	2	2	2	0	0

Table 1: Number of tanks used for combinations of benzopyrene concentration and fish age.

The analyses in Part I are all based on the tank average liver weights, computed by averaging the measured liver weight of all surviving fish in the tank. Throughout Part I,

$$\begin{aligned} Y_{ijk} \;=\; & \text{the average liver weight of fish of age } i \; (i = 1, 2) \\ & \text{exposed to concentration } j \; (j = 1, 2, \cdots 7) \\ & \text{in tank } k \; (k = 1, 2) \end{aligned}$$

where age 1 is adult and age 2 is juvenile and concentrations are indexed from smallest to largest.

A model that could be used to explore the joint effects of toxicant concentration and fish age on liver weight is:

$$\begin{aligned} \text{Model 1:} \quad Y_{ijk} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \epsilon_{ijk} = \mu_{ij} + \epsilon_{ijk} \\ \epsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma_1^2), \end{aligned}$$

for (i, j) pairs with positive sample sizes indicated in Table 1, and k = 1,2. Model 1 was fit to the data where fish age and concentration are categories. The default `R` full-rank parameterization ("set first to zero") was used. In this parameterization, the model matrix is made full rank by dropping every column corresponding to a parameter with any subscript of 1. `R` code and output for this part start on page 7.

Figure 1 is a plot of the tank average liver weights, identified by concentration and age of fish.

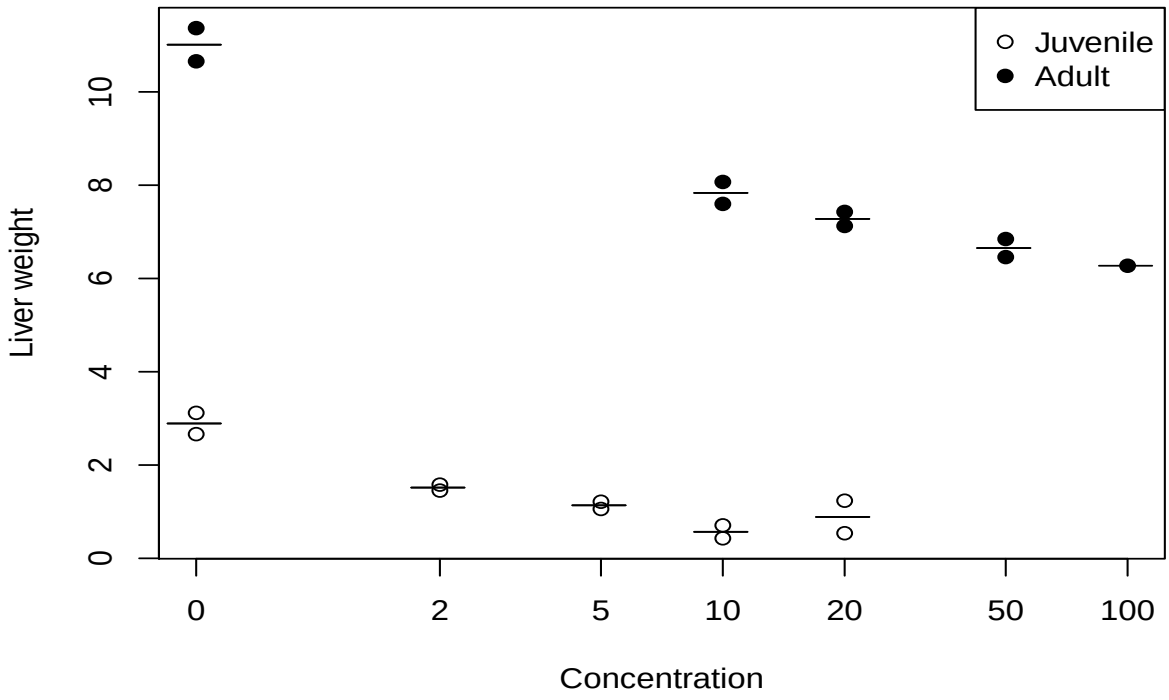


Figure 1: Plot of average liver weight values in each tank. Horizontal bars indicate the mean (across tanks) for each fish age and concentration.

1. Using the parameterization of Model 1, is $\mu_{13} - \mu_{23}$ estimable? Carefully explain why or why not.

2. Estimate $(\mu_{14}+\mu_{24})/2-(\mu_{11}+\mu_{21})/2$, the difference in expected response between the concentrations of 10 $\mu g/l$ and 0 $\mu g/l$, averaged over the two ages of fish.

3. Calculate a 95% confidence interval for the difference estimated in question 2.

4. Construct an appropriate hypothesis test of H_0: $\mu_{14} - \mu_{24} = \mu_{11} - \mu_{21}$. Report your test statistic, its distribution under H_0, and bound the p-value as closely as possible using the appropriate table of critical values.

5. Describe in words the hypothesis tested in question 4. Does the result of this test influence the interpretion of the estimate in question 2? Briefly explain.

A second model for the tank means is

$$\begin{aligned} \text{Model 2:} \quad Y_{ijk} &= \mu^*_{ij} + \epsilon_{ijk} \\ \mu^*_{ij} &= \begin{cases} \tau_i & C_{ijk} = 0 \\ \alpha_i + \beta_i \log C_{ijk} & C_{ijk} > 0 \end{cases} \\ \epsilon_{ijk} &\overset{iid}{\sim} N(0, \sigma^2_2), \end{aligned}$$

where C_{ijk} is the toxicant concentration in tank ijk.

Model 2 was fit to the data using the model matrix with columns shown in Table 2. To save space, only ten rows of the $\boldsymbol{X}$ matrix are shown, one for each observed age and concentration. `R` code and output from fitting this model begin on page 9. Table 2 corresponds to `fishmeanx` in the `R` code.

Group		Columns of model matrix					
Age	Conc.	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
A	0	1	0	0	0	0	0
A	10	0	1	2.303	0	0	0
A	20	0	1	2.996	0	0	0
A	50	0	1	3.912	0	0	0
A	100	0	1	4.605	0	0	0
J	0	0	0	0	1	0	0
J	2	0	0	0	0	1	0.693
J	5	0	0	0	0	1	1.609
J	10	0	0	0	0	1	2.303
J	20	0	0	0	0	1	2.996

Table 2: Some of the rows of the model matrix used to fit Model 2.

6. Describe in words what is quantified by $\tau_1 - \alpha_1 - \tau_2 + \alpha_2$.

7. Is $\mu^*_{13} - \mu^*_{23}$, i.e., the difference in mean liver weight between juvenile and adult fish exposed to 5 $\mu g/l$, estimable? Carefully explain why or why not.

8. The investigators are interested in the hypothesis

$$\mathrm{H}_0 : \alpha_1 = \alpha_2 \text{ and } \beta_1 = \beta_2.$$

Write out, but do not compute, an appropriate test statistic. Write out any matrices that do not depend on the responses (Y_{ijk})

9. Formally test the null hypothesis that Model 2 is a correct specification of the mean structure.

10. Carefully derive the distribution of your test statistic in question 9 under the null hypothesis that Model 2 is a correct specification of the mean structure. Include statements about symmetric and/or idempotent matrices where relevant to the derivation.

Part II

The analyses in Part II are all based on individual fish liver weights. Throughout Part II,

$$\begin{aligned} Y_{ijkl} \;=\; & \text{the individual liver weight of fish } l \; (l = 1, 2, \cdots, n_{ijk}) \\ & \text{of age } i \; (i = 1, 2) \\ & \text{exposed to concentration } j \; (j = 1, 2, \cdots 7) \\ & \text{in tank } k \; (k = 1, 2), \end{aligned}$$

where age 1 is adult and age 2 is juvenile and concentrations are indexed from smallest to largest. n_{ijk} is the number of surviving fish in tank ijk. n_{ijk} is not the same number for all tanks.

One possible model for individual liver weight is:

$$\begin{aligned} \text{Model 3:} \qquad Y_{ijkl} &= \mu + \delta_{ij} + \tau_{ijk} + \epsilon_{ijkl} \\ \tau_{ijk} &\overset{iid}{\sim} N(0, \sigma_3^2) \\ \epsilon_{ijkl} &\overset{iid}{\sim} N(0, \sigma_4^2), \end{aligned}$$

where μ and δ_{ij} are fixed parameters and all τ's and ϵ's are independent.

A similar model with fixed tank effects is:

$$\begin{aligned} \text{Model 4:} \qquad Y_{ijkl} &= \mu + \delta_{ij} + \tau_{ijk} + \epsilon_{ijkl} \\ \epsilon_{ijkl} &\overset{iid}{\sim} N(0, \sigma_4^2), \end{aligned}$$

where μ and δ_{ij} are fixed parameters as in Model 3 and τ_{ijk} is a fixed parameter associated with tank ijk.

`R` code and output from fitting Models 3 and 4 starts on page 11.

11. In the context of Model 3, test whether the observations, Y_{ijkl}, are independent. Results from fitting Model 4 may be helpful.

12. Consider tank effects to be random (Model 3). Calculate an appropriate standard error of $\hat{\mu} + \hat{\delta}_{14}$, i.e. the average liver weight of adult fish exposed to 10 μg/l toxicant. There are $n_{14k} = 10$ fish in each of the 2 tanks for this combination of age and concentration.

13. Explain how REML estimation of σ_3^2 and σ_4^2 in Model 3 differs from ML estimation.

Part III

The investigators also want to understand the effect of benzopyrene on mortality of adult fish. At the start of the study, each tank contained 10 fish. The number of surviving fish was recorded after 60 days of exposure to the toxicant. The data set named `a` contains information for each tank: `age` is the fish age (A), `conc` is the benzopyrene concentration, `survive` is the number of surviving fish, and `logconc` is the natural logarithm of the concentration.

```
a
  age conc survive  logconc
1   A   10      10 2.302585
2   A   10      10 2.302585
3   A   20      10 2.995732
4   A   20      10 2.995732
5   A   50       2 3.912023
6   A   50       8 3.912023
7   A  100       4 4.605170
8   A  100       0 4.605170
```

14. `R` code to fit one possible model (Model 5) to the mortality data for adult fish is:

```
a$sd <- cbind(a$survive,10-a$survive)
model.5 <- glm(sd~logconc, data=a, family='binomial')
```

Write out the statistical model being fit by this `R` code.

15. Figure 2 plots deviance residuals versus predicted values from fitting Model 5. Duplicate points are jittered in both directions. Using this information and all else you have been told about the study design and data collection, do the model assumptions in question 14 seem appropriate?

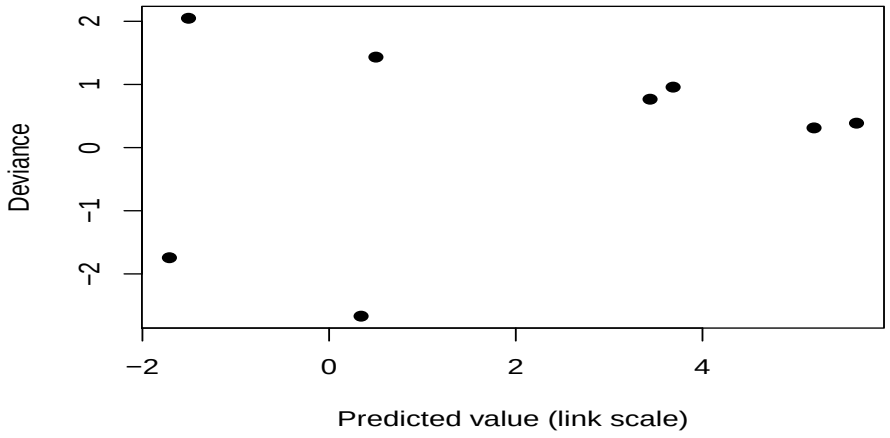


Figure 2: Predicted values and deviance residuals for each tank of adult fish.

Part IV

Whether or not a fish exposed to toxicant lives or dies depends partly on the amount of damage to its liver. That means that a fish with an unusually small liver is more likely to die than a fish with a large liver. Liver weight can only be recorded for the fish alive at the end of the study. The number of surviving individuals in each of the two tanks used for each fish age and concentration is given in Table 3.

		Concentration, μg/l						
		0	2	5	10	20	50	100
Fish Age	Adult	10, 10			10, 10	10, 10	2, 8	4, 0
	Juvenile	10, 10	10, 10	10, 10	10, 10	2, 6		

Table 3: Numbers of surviving fish per tank for each combination of benzopyrene concentration and fish age. Blanks in the table indicate combinations of concentration and fish age that were not included in the study.

16. What consequences, if any, does an association between liver weight and survival have for your interpretation of differences in mean liver weight analyzed in Part I? Explain.

17. Propose a model for the tank mean data analyzed in Part I that allows you to explore more carefully whether survival is an issue when studying the relationship between toxicant concentration and liver weight. Define any new variables you use in your model.

R code and output for questions in Part I

```
# R code for Model 1.

head(fishmean)
model.1 <- lm(meanwt ~ age + conc + age:conc, data=fishmean)
logLik(model.1)
summary(model.1)
anova(model.1)
vcov(model.1)

# R code for Model 2.

head(fishmeanx)
model.2 <- lm(meanwt ~ -1 + X1 + X2 + X3 + X4 + X5 + X6, data=fishmeanx)
logLik(model.2)
summary(model.2)
anova(model.2)
vcov(model.2)

------------------------------------------------------------
# Output for Model 1.

> head(fishmean)
  tank age conc     meanwt
1    1   J    0  3.1191781
2   10   J   20  0.5383345
3   11   A    0 11.3678893
4   12   A    0 10.6562999
5   13   A   10  7.5991597
6   14   A   10  8.0699889

> model.1 <- lm(meanwt ~ age + conc + age:conc, data=fishmean)
> summary(model.1)

Call:
lm(formula = meanwt ~ age + conc + age:conc, data = fishmean)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3558 -0.1722  0.0000  0.1722  0.3558

Coefficients: (4 not defined because of singularities)
```

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   11.0121     0.2225  49.503 2.81e-12 ***
ageJ          -8.1211     0.3146 -25.814 9.47e-10 ***
conc2         -1.3740     0.3146  -4.368 0.001803 **
conc5         -1.7545     0.3146  -5.577 0.000344 ***
conc10        -3.1775     0.3146 -10.100 3.29e-06 ***
conc20        -3.7351     0.3146 -11.873 8.43e-07 ***
conc50        -4.3597     0.3146 -13.858 2.24e-07 ***
conc100       -4.7390     0.3853 -12.300 6.24e-07 ***
ageJ:conc2         NA         NA      NA       NA
ageJ:conc5         NA         NA      NA       NA
ageJ:conc10    0.8546     0.4449   1.921 0.086925 .
ageJ:conc20    1.7312     0.4449   3.891 0.003668 **
ageJ:conc50        NA         NA      NA       NA
ageJ:conc100       NA         NA      NA       NA
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3146 on 9 degrees of freedom
Multiple R-squared: 0.9963,Adjusted R-squared: 0.9925
F-statistic: 266.6 on 9 and 9 DF,  p-value: 7.805e-10

> logLik(model.1)
'log Lik.' 2.111735 (df=11)

> anova(model.1)
Analysis of Variance Table

Response: meanwt
          Df  Sum Sq Mean Sq   F value    Pr(>F)
age        1 205.126 205.126 2072.6325 5.954e-12 ***
conc       6  30.880   5.147   52.0036 1.765e-06 ***
age:conc   2   1.499   0.749    7.5712   0.01179 *
Residuals  9   0.891   0.099
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> vcov(model.1)
              (Intercept)       ageJ         conc2         conc5        conc10
(Intercept)  4.948440e-02 -0.0494844 -2.193659e-18 -2.997755e-18 -4.948440e-02
ageJ        -4.948440e-02  0.0989688 -4.948440e-02 -4.948440e-02  4.948440e-02
conc2       -2.193659e-18 -0.0494844  9.896880e-02  4.948440e-02  1.055585e-18
```

```
conc5       -2.997755e-18 -0.0494844  4.948440e-02  9.896880e-02  1.010116e-17
conc10      -4.948440e-02  0.0494844  1.055585e-18  1.010116e-17  9.896880e-02
conc20      -4.948440e-02  0.0494844  4.387318e-18  5.995510e-18  4.948440e-02
conc50      -4.948440e-02  0.0494844 -1.691351e-18 -8.865844e-19  4.948440e-02
conc100     -4.948440e-02  0.0494844  5.841263e-19  1.388893e-18  4.948440e-02
ageJ:conc10  4.948440e-02 -0.0989688  4.948440e-02  4.948440e-02 -9.896880e-02
ageJ:conc20  4.948440e-02 -0.0989688  4.948440e-02  4.948440e-02 -4.948440e-02
                    conc20        conc50       conc100 ageJ:conc10 ageJ:conc20
(Intercept) -4.948440e-02 -4.948440e-02 -4.948440e-02   0.0494844   0.0494844
ageJ         4.948440e-02  4.948440e-02  4.948440e-02  -0.0989688  -0.0989688
conc2        4.387318e-18 -1.691351e-18  5.841263e-19   0.0494844   0.0494844
conc5        5.995510e-18 -8.865844e-19  1.388893e-18   0.0494844   0.0494844
conc10       4.948440e-02  4.948440e-02  4.948440e-02  -0.0989688  -0.0494844
conc20       9.896880e-02  4.948440e-02  4.948440e-02  -0.0494844  -0.0989688
conc50       4.948440e-02  9.896880e-02  4.948440e-02  -0.0494844  -0.0494844
conc100      4.948440e-02  4.948440e-02  1.484532e-01  -0.0494844  -0.0494844
ageJ:conc10 -4.948440e-02 -4.948440e-02 -4.948440e-02   0.1979376   0.0989688
ageJ:conc20 -9.896880e-02 -4.948440e-02 -4.948440e-02   0.0989688   0.1979376
```

```
# Output for model 2.

> head(fishmeanx)
  X1 X2       X3 X4 X5       X6     meanwt
1  1  0 0.000000  0  0 0.000000  3.1191781
2  0  1 2.995732  0  0 0.000000  0.5383345
3  0  0 0.000000  1  0 0.000000 11.3678893
4  0  0 0.000000  1  0 0.000000 10.6562999
5  0  0 0.000000  0  1 2.302585  7.5991597
6  0  0 0.000000  0  1 2.302585  8.0699889

> model.2 <- lm(meanwt ~ -1 + X1 + X2 + X3 + X4 + X5 + X6, data=fishmeanx)
> summary(model.2)

Call:
lm(formula = meanwt ~ -1 + X1 + X2 + X3 + X4 + X5 + X6, data = fishmeanx)

Residuals:
     Min       1Q   Median       3Q      Max
-0.46477 -0.19446  0.02647  0.15808  0.57136

Coefficients:
```

```
   Estimate Std. Error t value Pr(>|t|)
X1   2.8910     0.2176  13.283 6.12e-09 ***
X2   1.6562     0.2660   6.227 3.08e-05 ***
X3  -0.3310     0.1277  -2.592 0.022356 *
X4  11.0121     0.2176  50.597 2.57e-16 ***
X5   9.3889     0.4849  19.364 5.71e-11 ***
X6  -0.6914     0.1431  -4.832 0.000328 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.3078 on 13 degrees of freedom
Multiple R-squared: 0.998,Adjusted R-squared: 0.9971
F-statistic:  1099 on 6 and 13 DF,  \emph{p}-value: < 2.2e-16

> logLik(model.2)
'log Lik.' -0.9666095 (df=7)

> anova(model.2)
Analysis of Variance Table

Response: meanwt
          Df Sum Sq Mean Sq   F value    Pr(>F)
X1         1  16.72   16.72  176.4467 6.122e-09 ***
X2         1   8.44    8.44   89.1010 3.499e-07 ***
X3         1   0.64    0.64    6.7168 0.0223559 *
X4         1 242.53  242.53 2560.0309 2.572e-16 ***
X5         1 354.31  354.31 3739.8486 < 2.2e-16 ***
X6         1   2.21    2.21   23.3485 0.0003275 ***
Residuals 13   1.23    0.09
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> vcov(model.2)
           X1          X2          X3         X4          X5          X6
X1 0.04736905  0.00000000  0.00000000 0.00000000  0.00000000  0.00000000
X2 0.00000000  0.07074896 -0.03099984 0.00000000  0.00000000  0.00000000
X3 0.00000000 -0.03099984  0.01631377 0.00000000  0.00000000  0.00000000
X4 0.00000000  0.00000000  0.00000000 0.04736905  0.00000000  0.00000000
X5 0.00000000  0.00000000  0.00000000 0.00000000  0.23509459 -0.06735577
X6 0.00000000  0.00000000  0.00000000 0.00000000 -0.06735577  0.02047657
```

R code and output for questions in Part II

```
library(lme4)

# R code for Model 3.

head(fish)
model.3 <- lmer(wt ~ ageconc + (1|tank), data=fish)
summary(model.3)
logLik(model.3)
anova(model.3)

# R code for Model 4.

model.4 <- lm(wt ~ ageconc + tank, data=fish)
summary(model.4)
logLik(model.4)
anova(model.4)

-----------------------------------------------------------------------
# Output for Model 3.

> head(fish)
  tank       wt ageconc
1    1 3.589478     J/0
2    1 3.531171     J/0
3    1 3.233566     J/0
4    1 3.204037     J/0
5    1 3.153545     J/0
6    1 3.146007     J/0


model.3 <- lmer(wt ~ ageconc + (1|tank), data=fish)
summary(model.3)

Linear mixed model fit by REML
Formula: wt ~ ageconc + (1 | tank)
   Data: fish
   AIC   BIC logLik deviance REMLdev
 113.7 150.8 -44.85    78.06    89.7
Random effects:
 Groups   Name        Variance Std.Dev.
 tank     (Intercept) 0.083917 0.28968
```

```
 Residual             0.077557 0.27849
Number of obs: 162, groups: tank, 19

Fixed effects:
             Estimate Std. Error t value
(Intercept)   11.0121     0.2141   51.44
ageconcA/10   -3.1775     0.3028  -10.49
ageconcA/100  -4.7390     0.3862  -12.27
ageconcA/20   -3.7351     0.3028  -12.34
ageconcA/50   -4.3857     0.3145  -13.94
ageconcJ/0    -8.1211     0.3028  -26.82
ageconcJ/10  -10.4439     0.3028  -34.50
ageconcJ/2    -9.4951     0.3028  -31.36
ageconcJ/20  -10.1660     0.3162  -32.16
ageconcJ/5    -9.8755     0.3028  -32.62

> logLik(model.3)
'log Lik.' -44.85043 (df=12)

> anova(model.3)
Analysis of Variance Table
        Df Sum Sq Mean Sq F value
ageconc  9 196.31  21.812  281.23

------------------------------------------------------------
# Output for Model 4.

> model.4 <- lm(wt ~ ageconc + tank, data=fish)
> summary(model.4)
Call:
lm(formula = wt ~ ageconc + tank, data = fish)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89925 -0.20064 -0.01629  0.19700  0.59240

Coefficients: (9 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.65630    0.08802 121.068  < 2e-16 ***
ageconcA/10   -2.58631    0.12448 -20.777  < 2e-16 ***
ageconcA/100  -4.38321    0.16467 -26.618  < 2e-16 ***
ageconcA/20   -3.22854    0.12448 -25.937  < 2e-16 ***
```

```
ageconcA/50   -4.19760    0.13203 -31.793  < 2e-16 ***
ageconcJ/0    -7.53712    0.12448 -60.550  < 2e-16 ***
ageconcJ/10  -10.22706    0.12448 -82.160  < 2e-16 ***
ageconcJ/2    -9.07557    0.12448 -72.909  < 2e-16 ***
ageconcJ/20  -10.11797    0.14373 -70.394  < 2e-16 ***
ageconcJ/5    -9.44213    0.12448 -75.854  < 2e-16 ***
<part of output deleted>
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2783 on 143 degrees of freedom
Multiple R-squared: 0.9949,Adjusted R-squared: 0.9943
F-statistic:  1552 on 18 and 143 DF,  p-value: < 2.2e-16

> logLik(model.4)
'log Lik.' -12.57959 (df=20)

> anova(model.4)
Analysis of Variance Table

Response: wt
           Df  Sum Sq Mean Sq   F value    Pr(>F)
ageconc     9 2157.86 239.763 3094.7768 < 2.2e-16 ***
tank        9    6.69   0.744    9.5999 2.221e-11 ***
Residuals 143   11.08   0.077
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Part I

1. No. $\mu_{13} - \mu_{23} = (\mu + \alpha_1 + \beta_3 + \alpha\beta_{13}) - (\mu + \alpha_2 + \beta_3 + \alpha\beta_{23})$. This can not be expressed as a linear combination of rows of the non-full rank model matrix since the column of the non-full rank model matrix for $\alpha\beta_{13}$ is always zero. There are no observations that provide information about $\alpha\beta_{13}$.

2. $(\mu_{14} + \mu_{24})/2 - (\mu_{11} + \mu_{21})/2 = ((\mu + \alpha_1 + \beta_4 + \alpha\beta_{14}) + (\mu + \alpha_2 + \beta_4 + \alpha\beta_{24}))/2 - ((\mu + \alpha_1 + \beta_1 + \alpha\beta_{11}) + (\mu + \alpha_2 + \beta_1 + \alpha\beta_{21}))/2 = \beta_4 - \beta_1 + (\alpha\beta_{14} + \alpha\beta_{24} - \alpha\beta_{11} - \alpha\beta_{21})/2$
Under R's set first to zero parameterization, β_1, $\alpha\beta_{14}$, $\alpha\beta_{11}$, and $\alpha\beta_{21}$ are all 0. The quantity of interest is $\beta_4 + \alpha\beta_{24}/2$, which is estimated as -3.1775 + 0.8546/2 = -2.750.

3. The variance of the quantity of interest is $\text{Var}(\hat{\beta}_2)$ + $\text{Var}(\widehat{\alpha\beta}_{24}/4)$ - 2 $\text{Cov}(\hat{\beta}_2, \widehat{\alpha\beta}_{24}/2)$ = 0.09896 + 0.19794/4 + 2(-0.0989688)/2 = 0.04948. The Error SS has 9 degrees of freedom, so the t quantile for a 95% confidence interval is $t_{0.975,9} = 2.262$. The answer is $(-2.750 - 2.262\sqrt{0.04948},\ -2.750 + 2.262\sqrt{0.04948}) = (-3.25, -2.25)$.

4. A similar process gives $(\mu_{14} - \mu_{24}) - (\mu_{11} - \mu_{21}) = -\alpha\beta_{24}$. The t statistic for the the test of $\alpha\beta_{14} = 0$ is obtainable from the summary output. t = 1.921. This has a t distribution with 9 d.f. under Ho: $\alpha\beta_{24} = 0$. The p-value is 0.087.

5. The test in question 4 evaluates whether a concentration of $10\mu g/l$ has the same effect in adults and juveniles, where effect is measured as the difference from the control treatment. That is, is there an interaction between age and concentration, for the control and 10 $\mu g/l$ treatments.
The quantity in question 2 is the difference between $10\mu g/l$ and the control, averaged over adults and juveniles. If you are willing to believe there is no interaction, the average difference can be interpreted as the simple effect at each age.

6. $(\tau_1 - \alpha_1) - (\tau_2 - \alpha_2)$ quantifies whether the effect of a concentration of 1 $\mu g/l$ of toxicant (i.e., at $\log C_{ijk} = 0$) is the same in adults and juveniles, where effect is measured as the difference from the control.

7. Yes. A quantity is estimable if it can be expressed as a linear combination of rows of $\boldsymbol{X\beta}$. The quantity desired is $(\alpha_1 + 1.609\beta_1) - (\alpha_2 + 1.609\beta_2)$. β_1 and β_2 are linear combinations of rows of $\boldsymbol{X\beta}$ (e.g. corresponding to juvenile, 5 $\mu g/l$ and juvenile, 10 $\mu g/l$ or adult, 10 $\mu g/l$ and adult, 20 $\mu g/l$). α_1 and α_2 are one row of the $\boldsymbol{X\beta}$ matrix minus the appropriate β times the value of $X3$. Hence, the desired quantity is a linear combination of rows of $\boldsymbol{X\beta}$.

8. The hypothesis can be tested in various ways. The test statistic that can be computed from the information available is a F statistic to test $\boldsymbol{C\beta = d}$. The parameter vector,

$\boldsymbol{\beta}$, is $[\tau_1\ \alpha_1\ \beta_1\ \tau_2\ \alpha_2\ \beta_2]$, so the appropriate $\boldsymbol{C}$ matrix is

$$\boldsymbol{C} = \begin{bmatrix} 0 & 1 & 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{bmatrix}$$

and $\boldsymbol{d} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. The test statistic is then

$$F = (\boldsymbol{C}\hat{\boldsymbol{\beta}})' \left(\boldsymbol{C}'\hat{\boldsymbol{\Sigma}}\boldsymbol{C}\right)^{-1} (\boldsymbol{C}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated variance covariance matrix of $\hat{\boldsymbol{\beta}}$, where $\boldsymbol{\beta}$ is the parameter vector defined above.

9. Since the data set includes multiple observations at the same experimental conditions, you can use an ANOVA lack of fit test comparing Model 1 (separate means for each unique set of experimental conditions) and Model 2 (the proposed regression model).

Source	df	SS	MS
Lack of Fit	4	0.341	0.0852
Model 1	9	0.891	0.099
Model 2	13	1.232	

The F statistic is $F = \frac{0.0852}{0.099} = 0.861$. The p-value is > 0.5.

10. The outline of the derivation is to show:
1) Error SS / $\sigma^2 \sim$ Chi-square with 9 df
2) Lack of Fit SS / $\sigma^2 \sim$ Chi-square with 4 df
3) Lack of Fit SS and Error SS are independent
4) so ratio is central F with 4, 9 df.

The details of each step:
Define $\boldsymbol{X_1}$ as the model matrix for fitting model 1 (separate means), $\boldsymbol{X_2}$ as the model matrix for fitting model 2 (proposed regression), $\boldsymbol{P_{X_1}} = \boldsymbol{X_1}(\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'}$ is the projection matrix into the column space of $\boldsymbol{X_1}$, and $\boldsymbol{P_{X_2}} = \boldsymbol{X_2}(\boldsymbol{X_2'X_2})^{-1}\boldsymbol{X_2'}$ is the projection matrix into the column space of $\boldsymbol{X_2}$.
We assume $\boldsymbol{Y} \sim N(\boldsymbol{X_1\beta}, \sigma^2)$. Under H_0: $\boldsymbol{Y} \sim N(\boldsymbol{X_2\beta}, \sigma^2)$ 1) The error SS/σ^2 = $\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{P_{X_1}})\boldsymbol{Y}'/\sigma^2$, which is of the form $\boldsymbol{W}'\boldsymbol{AW}$ for $\boldsymbol{A} = (\boldsymbol{I} - \boldsymbol{P_{X_1}})/\sigma^2$. $\boldsymbol{A\Sigma} = (\boldsymbol{I} - \boldsymbol{P_{X_1}})$, which is idempotent. Hence error SS/$\sigma^2 \sim$ Chi-square with df = $tr(\boldsymbol{I} - \boldsymbol{P_{X_1}})$ = $N -$ rank $\boldsymbol{P_{X_1}}$ = 19 - 10 = 9
2) The lack of fit SS / σ^2 = $\boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{P_{X_2}})\boldsymbol{Y}'/\sigma^2 - \boldsymbol{Y}(\boldsymbol{I} - \boldsymbol{P_{X_1}})\boldsymbol{Y}'/\sigma^2 = \boldsymbol{Y}(\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_2}})\boldsymbol{Y}'/\sigma^2$. Argument similar to 1) gives lack of fit SS /$\sigma^2 \sim$ central Chi-square with df = $tr(\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_2}})$ = rank $\boldsymbol{P_{X_1}}$ − rank $\boldsymbol{P_{X_2}}$ = 10 - 6 = 4

3) The two SS are independent if $(\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_2}})\boldsymbol{\Sigma}\boldsymbol{P_{X_2}} = \mathbf{0}$.

$$\begin{aligned}(\boldsymbol{P_{X_1}} - \boldsymbol{P_{X_2}})\boldsymbol{\Sigma}\boldsymbol{P_{X_2}} &= \sigma^2(\boldsymbol{P_{X_1}}\boldsymbol{P_{X_2}} - \boldsymbol{P_{X_2}}\boldsymbol{P_{X_2}}) \\ &= \sigma^2(\boldsymbol{P_{X_2}} - \boldsymbol{P_{X_2}}) \\ &= \mathbf{0}\end{aligned}$$

4) Under H_0, the F statistic to test lack of fit in question 9 can be written as $F = (U_1/df_1)/(U_2/df_2)$ where U_1 has a central Chi-square distribution with df_1 degrees of freedom and U_2 has a central Chi-square distribution with df_2 degrees of freedom. Hence, under H_0, the test statistic has a central F distribution with 4, 9 degrees of freedom.

Part II

11. The observations on each individual fish are independent if and only if the variance component for tanks, σ_3^2, $= 0$ in model 3. This hypothesis is equivalent to the hypothesis that all tanks have the same mean (model 4). This hypothesis can be easily tested with the information available. Using the ANOVA table for model 4, the F statistic for tanks after fitting treatment means is 9.60. The p-value is < 0.0001.

12. The estimated variance of the mean of 2 tanks, with 10 fish per tank, is

$$\frac{\hat{\sigma}_3^2}{2} + \frac{\hat{\sigma}_4^2}{2 \times 10} = \frac{0.08392}{2} + \frac{0.07756}{2 \times 10} = 0.0458.$$

The standard error is $\sqrt{0.0458} = 0.21$.

13. ML estimation maximizes the log likelihood for a multivariate normal distribution for the entire vector of observations, $\boldsymbol{Y}$. REML estimation maximizes the log likelihood for a multivariate normal distribution for the a vector of residuals, $\boldsymbol{Y}^*$, where $\boldsymbol{Y}^*$ is $N - k$ linearly independent rows of $\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{Y}$.

Part III

14.

$$\begin{aligned}Y_{ij} &\sim \text{Binomial}(10, \pi_i) \\ \log\left(\frac{\pi_i}{1-\pi_i}\right) &= \beta_0 + \beta_1 X_i\end{aligned}$$

Y_{ij} is the number of surviving fish in tank j and concentration level i. X_i is the log of the toxicant concentration at level i. β_0 and β_1 are the intercept and slope of the logistic equation relating survival to log toxicant concentration.

15. Because there are so few tanks, it is hard to tell much about lack of fit to the proposed regression model. The model does assume that the two replicate tanks at each concentration have the same π_i. This is equivalent to the assumption that survival of each fish is an independent Bernoulli random variable. If the model fits, the deviance residuals are approximately $\sim N(0,1)$. With 9 values from $N(0,1)$, you would expect few values close to -2 or +2. In fact, four of the nine deviance residuals are near +2 or -2. These are from the four tanks with low survival. The large spread in the deviance residuals suggests there is substantial tank-tank variation within the same treatment. Hence, the assumption of independence of individual fish is suspect.

Part IV

16. Yes, it is a potential problem. The assumption in all analyses in parts 1 and 2 is that observations are missing completely at random. If survival is a function of liver weight, the missing observations (dead fish) are not missing completely at random. This biases the estimates of treatment effects.

17. One simple idea would be to assess whether the number of deaths is associated with the mean liver weight for a given fish age and toxicant concentration. The form of this relationship is not clear, so a simple approach is to assume a linear relationship. The resulting model for tank mean liver weight would be:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \alpha\beta_{ij} + \gamma N_{ijk} + \epsilon_{ijk},$$

where N_{ijk} is the number of deaths in tank ijk and all other quantities are defined as in model 1.
Other models are also very acceptable.

Problem Background and Data

Retail grocers are aware that effects of many factors determine the quantities of various goods that they will sell. Untangling these factors is a major problem that most grocers approach through simple analyses of available data and intuition. A part of the reason for this approach is that formulation of appropriate statistical models is difficult. The problem to be addressed in this question is development of a statistical model for sales of one good, green beans, as related to one covariate, price, for a grocery chain in the Midwestern United States.

The available data consist of records of daily price and number of units sold for 179 grocery stores over periods ranging from roughly October 2005 through September 2007 (about 725 days). The time period of data availability varies from store to store, but most stores have somewhere between 700 and 720 days of data recorded (the minimum number of days of data is 165, but the next lowest is 400 and the first quartile is 705). Within each store, days recorded are consecutive and there are no gaps in the temporal sequences of data for individual stores.

Scatterplots of units sold versus price are presented for four individual stores in Figure 1. Note that although in this figure, and others to follow, the identities of these four particular stores are given, that information is irrelevant to this question, as we will be attempting to formulate models appropriate for describing the relation between sales and price for all 179 stores. These four stores were selected for examination because they exhibit data patterns that seem to be generally present in the larger data set. Several aspects of the data are evident in the scatterplots of Figure 1. First, the number of unique prices for an individual store is small, ranging from about 6 to 12. These prices have values between 0.25 and 0.80. Units of price and amount sold were not provided with the data, but the values in Figure 1 might suggest price in dollars for units sold in cans of green beans. Secondly, there appears to be an inverse relation between price and number of units sold, with variability being greater for low prices and higher numbers of units sold. One visual impression produced by the plots in Figure 1 is misleading, that being the number of observations that seem available at the various prices. This visual impression is an artifact of the variances and "overplotting" of points with a given price. For example, in the upper left panel of Figure 1 it appears that there may be more observations for some

of the lower prices than for higher prices, but this is not the case. For this store, there were 65 observations at price 0.40, but 387 observations at price 0.67, and this pattern is true for stores in general.

The data were collected over time, and plots of units sold over time are presented for our four example stores in Figure 2. Similarly, plots of price over time are presented in Figure 3. The plots of Figure 2 show that high sales were short-lived (i.e., did not last long over time) and occurred throughout the time sequences of data collection in these four stores. Prices, which are fixed by management, changed frequently as demonstrated by the plots of Figure 3, but lower prices tended to have shorter "runs" than higher prices. We might surmise from Figures 1-3 that lower prices are frequent but do not last long, and that the number of units of green beans sold responds to price reductions with spikes of high volume days at the time of low prices. There is also quite a bit of variability in sales volume of green beans at low prices, as evidenced by Figure 1.

Part I: Models for Individual Stores

We might begin the process of developing a model to relate sales of green beans to price over all 179 stores by first formulating models for these variables using data from individual stores. We will continue to use the four individual stores of Figures 1-3 as examples for this process. To this end, define response random variables $\{Y_i : i = 1, \ldots, n\}$ to be connected with the number of units of green beans sold on day i in an individual store. Also define $\{x_i : i = 1, \ldots, n\}$ as the corresponding prices. Note from Figure 1 that the response random variables are discrete with many small (0 and single digit) and a few quite large (200 or more) values. The data contain a large number of replicate values at many price values. Empirical probability functions are presented in Figure 4 for sales at the four example stores. These plots are presented for selected values of price at which many replicate observations were available; these prices were 0.63 for the store depicted in the lower right panel, 0.67 for the store depicted in the upper left panel, and 0.68 for the two remaining stores. The most notable feature of these plots is the exaggerated relative frequency for values of 0 units sold. Aside from this elevated frequency, the distributions might be described by a unimodal right skewed distribution for discrete non-negative random variables. Although evidence has not been presented for other levels of price, it appears that at least for the realizations of Figure 4, a reasonable model might be a "zero inflated Poisson," which results from mixing a Poisson distribution with a binary

distribution. The resulting distribution, applied to an individual random variable Y_i, has probability mass function, for $\lambda > 0$ and $0 < p < 1$,

$$f(y|p,\lambda) = \begin{cases} (1-p) + p\exp(-\lambda) & y = 0 \\ \frac{1}{y!} p\lambda^y \exp(-\lambda) & y = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

ANSWER QUESTIONS 1, 2, and 3 NOW.

(Questions begin after the section with Figures.)

For subsets of the random variables Y_i that all share a common value of price x_i, moment estimates of p and λ were computed for each of the four example stores. Scatterplots of these estimates $\hat{p}$ and $\hat{\lambda}$ against price are presented in Figure 5. These plots suggest that both the mixing probability p and the Poisson parameter λ are decreasing with price, at least in these example stores. For random variables $\{Y_i : i = 1, \ldots, n\}$ and price covariates $\{x_i : i = 1, \ldots, n\}$ for a given store, the exploratory plots presented so far suggest a model of the following form. Let the Y_i be independent with probability mass functions

$$f_i(y|p_i,\lambda_i) = \begin{cases} (1-p_i) + p_i\exp(-\lambda_i) & y = 0 \\ \frac{1}{y!} p_i\lambda_i^y \exp(-\lambda_i) & y = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

such that

$$p_i = g_p(x_i, \boldsymbol{\beta}) \qquad \text{and} \qquad \lambda_i = g_\lambda(x_i, \boldsymbol{\gamma}), \tag{3}$$

where $g_p(\cdot)$ and $g_\lambda(\cdot)$ are known functions and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are unknown parameter vectors.

ANSWER QUESTION 4 NOW.

After additional exploratory analysis that is not presented here, the following forms were selected for $g_p(\cdot)$ and $g_\lambda(\cdot)$ in (3).

$$\begin{aligned} p_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \\ \lambda_i &= \exp\left[\gamma_0 + \gamma_1 \log(x_i)\right]. \end{aligned} \tag{4}$$

The model of expressions (2) and (4) was fit to data from the four example stores using maximum likelihood estimation. Fitted relations are presented in Figure 6.

ANSWER QUESTIONS 5, 6, 7 and 8 NOW.

Recall that the observed data within each store were collected over sequences of hundreds of days. Autocorrelation plots for the number of units sold are presented in Figure 7 for the four example stores.

ANSWER QUESTION 9 NOW.

Part II: Models for Many Stores

Suppose that we have either determined that the model of expressions (2) and (4) is largely adequate for dealing with data from individual stores, or that we are at least willing to use that model structure as a starting point for developing a model that can deal with many stores simultaneously. Extend the notational system used to this point in the question as follows. Let $\{Y_{i,j} : i = 1, \ldots, n_j;\ j = 1, \ldots, S\}$ denote random variables connected with the number of units of green beans sold in store j on day i and let $\{x_{i,j} : i = 1, \ldots, n_j;\ j = 1, \ldots, S\}$ denote the corresponding prices; recall that here $S = 179$. Let the $Y_{i,j}$ be conditionally independent with probability mass functions

$$f_i(y|p_{i,j},\, \lambda_{i,j}) = \begin{cases} (1 - p_{i,j}) + p_{i,j} \exp(-\lambda_{i,j}) & y = 0 \\ \frac{1}{y!} p_{i,j} \lambda_{i,j}^{y} \exp(-\lambda_{i,j}) & y = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

Three techniques that present themselves for modeling data across multiple stores are as follows.

1. Model the $p_{i,j}$ and $\lambda_{i,j}$ as,

$$\begin{aligned} p_{i,j} &= \frac{\exp(\beta_0 + \beta_1 x_{i,j})}{1 + \exp(\beta_0 + \beta_1 x_{i,j})} + \epsilon_j, \\ \lambda_{i,j} &= \exp\left[\gamma_0 + \gamma_1 \log(x_{i,j})\right] + \xi_j, \end{aligned} \tag{6}$$

   where the ϵ_j are assumed to be independent and identically distributed according to some location-scale family of distributions (most likely normal) having expected value 0 and variance τ_p^2 and, similarly, the ξ_j are assumed to be independent and identically distributed according to some location-scale family of distributions having expected value 0 and variance τ_λ^2, and we further assume that the ϵ_j and ξ_j are independent, unless subsequent evidence indicates this would not be a good assumption to make.

2. Model the $p_{i,j}$ and $\lambda_{i,j}$ as,

$$\begin{aligned} p_{i,j} &= \frac{\exp(\beta_0 + \beta_1 x_{i,j} + \epsilon_j)}{1 + \exp(\beta_0 + \beta_1 x_{i,j} + \epsilon_j)}, \\ \lambda_{i,j} &= \exp\left[\gamma_0 + \gamma_1 \log(x_{i,j} + \xi_j)\right], \end{aligned} \tag{7}$$

where the ϵ_j are assumed to be independent and identically distributed according to some location-scale family of distributions (most likely normal) having expected value 0 and variance τ_p^2 and, similarly, the ξ_j are assumed to be independent and identically distributed according to some location-scale family of distributions having expected value 0 and variance τ_λ^2, and we further assume that the ϵ_j and ξ_j are independent, unless subsequent evidence indicates this would not be a good assumption to make.

3. Model the $p_{i,j}$ and $\lambda_{i,j}$ as,

$$\begin{aligned} p_{i,j} &= \frac{\exp(\beta_{0,j} + \beta_{1,j} x_{i,j})}{1 + \exp(\beta_{0,j} + \beta_{1,j} x_{i,j})}, \\ \lambda_{i,j} &= \exp\left[\gamma_{0,j} + \gamma_{1,j} \log(x_{i,j})\right], \end{aligned} \tag{8}$$

where $(\beta_{0,j}, \beta_{1,j}, \gamma_{0,j}, \gamma_{1,j})^T$ are independent and identically distributed according to some joint distribution F.

ANSWER QUESTION 10 NOW.

The basic model of expressions (2) and (3) was fit separately to data from each of the 179 stores using maximum likelihood estimation. Histograms of the resultant sets of estimated parameters are presented in Figure 8, and a scatterplot matrix of those estimates is presented in Figure 9.

We can also examine the ways that the model of (7) and that of (8) represent the relation between the $p_{i,j}$ and the covariate of price, x_{ij}. Figure 10 presents simulated values of $p_{i,j}$ from these models. Model (7) is shown in the upper panel using 5 values of ϵ_j chosen independently from a standard normal distribution. Model (8) is shown in the lower panel using 5 values of $\beta_{0,j}$ and $\beta_{1,j}$ randomly selected from the set of 179 fits to individual stores.

ANSWER QUESTIONS 11, 12, 13 and 14 NOW.

1 Figures

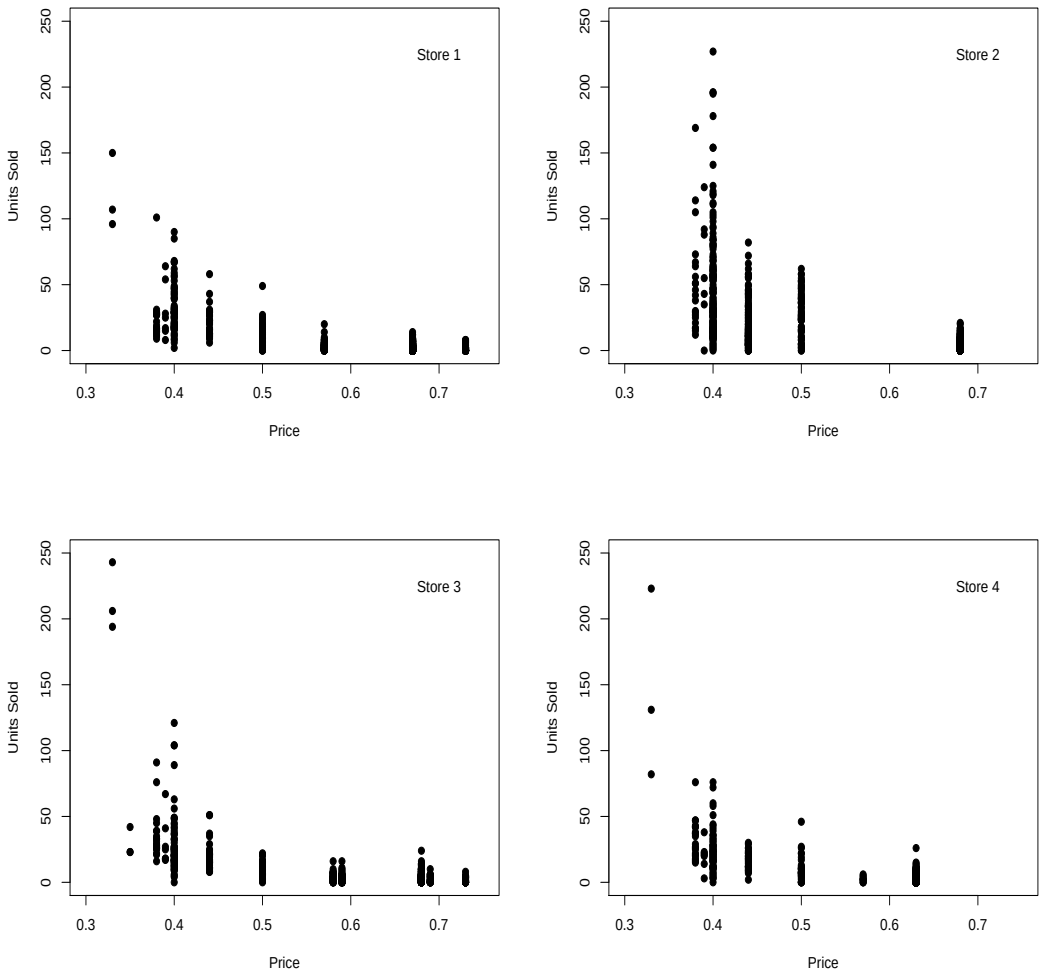


Figure 1: Scatterplots of units sold versus price for four individual stores.

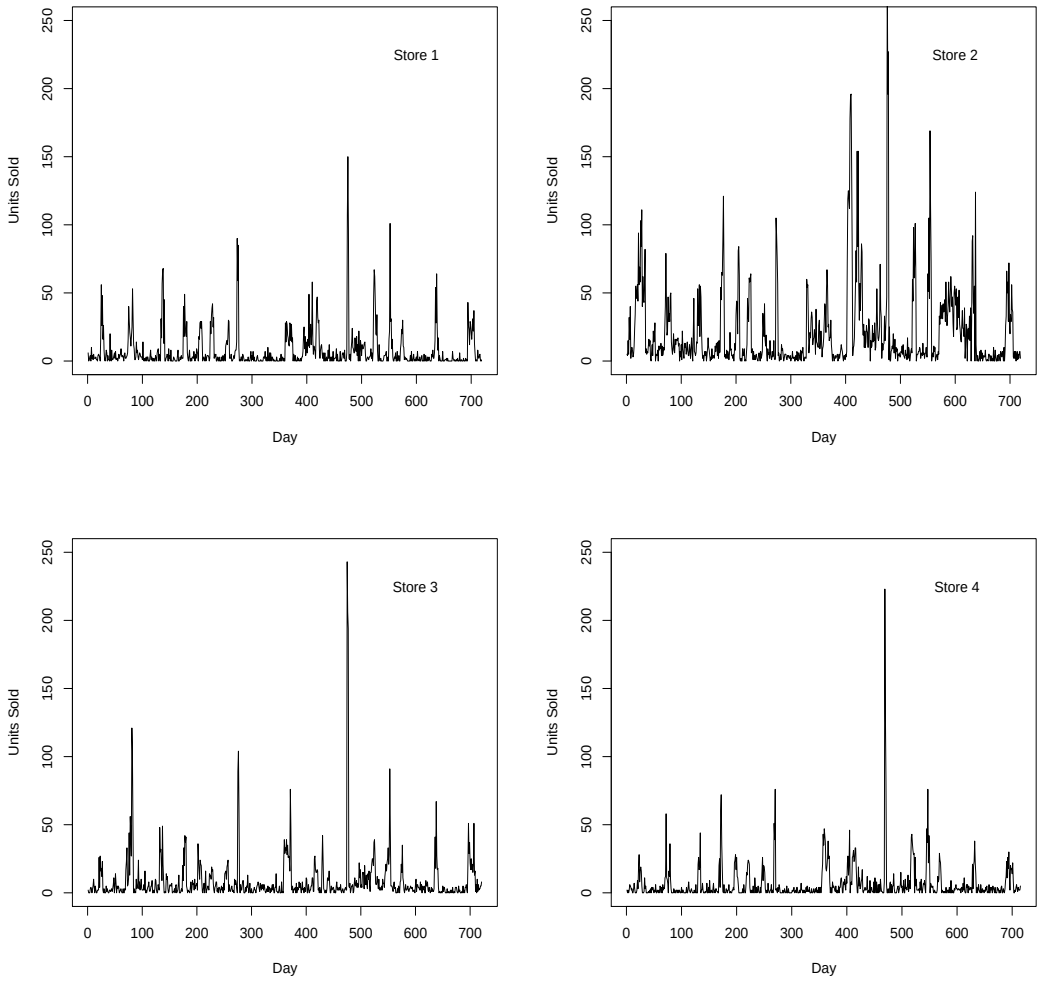


Figure 2: Time series plots of units sold for four individual stores.

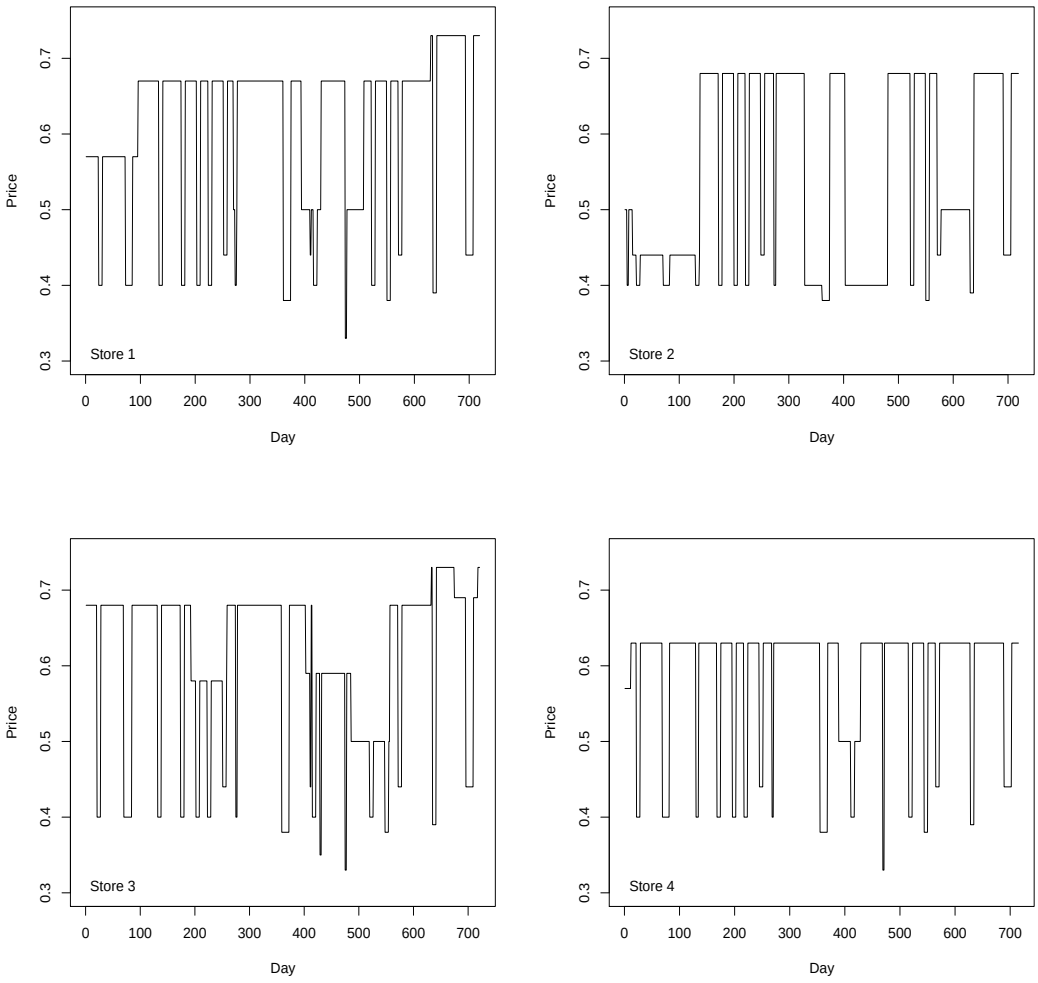


Figure 3: Time series plots of price for four individual stores.

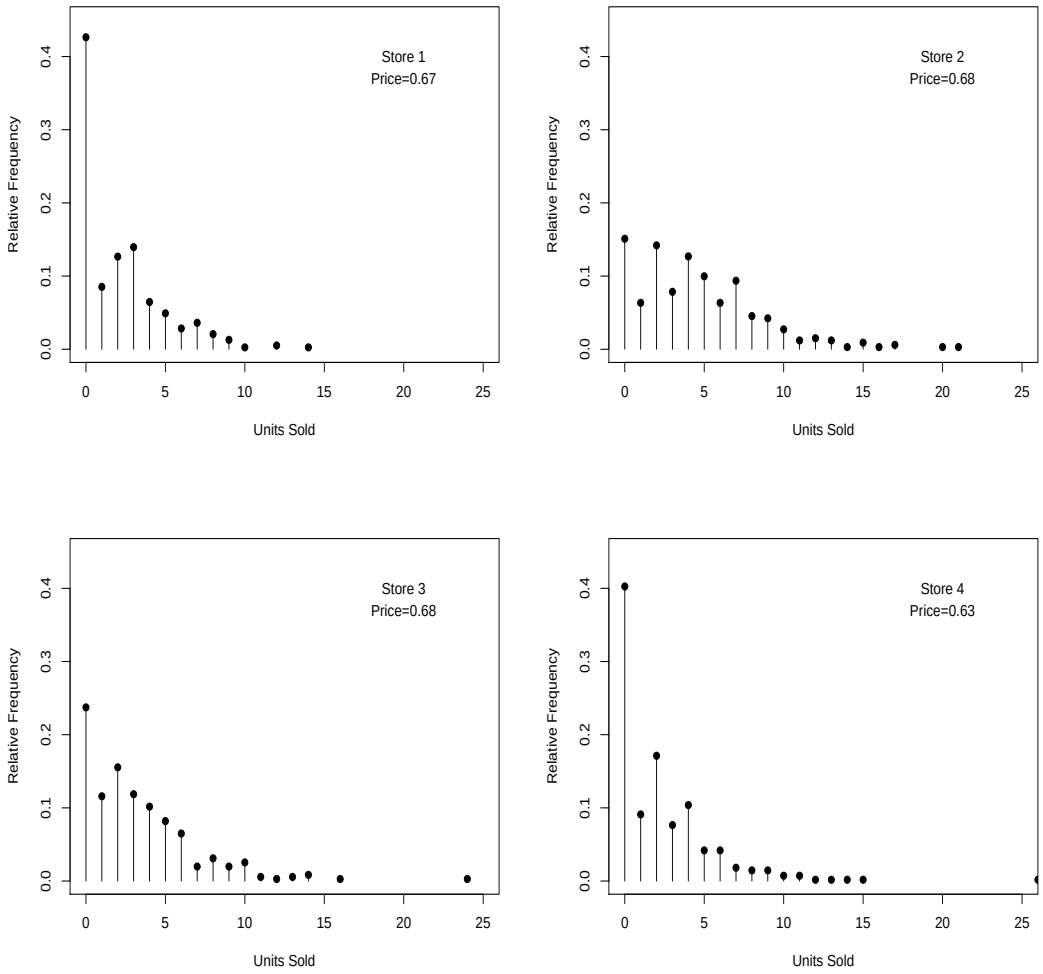


Figure 4: Empirical probability functions for sales at particular prices ranging from 0.63 to 0.68 for four individual stores.

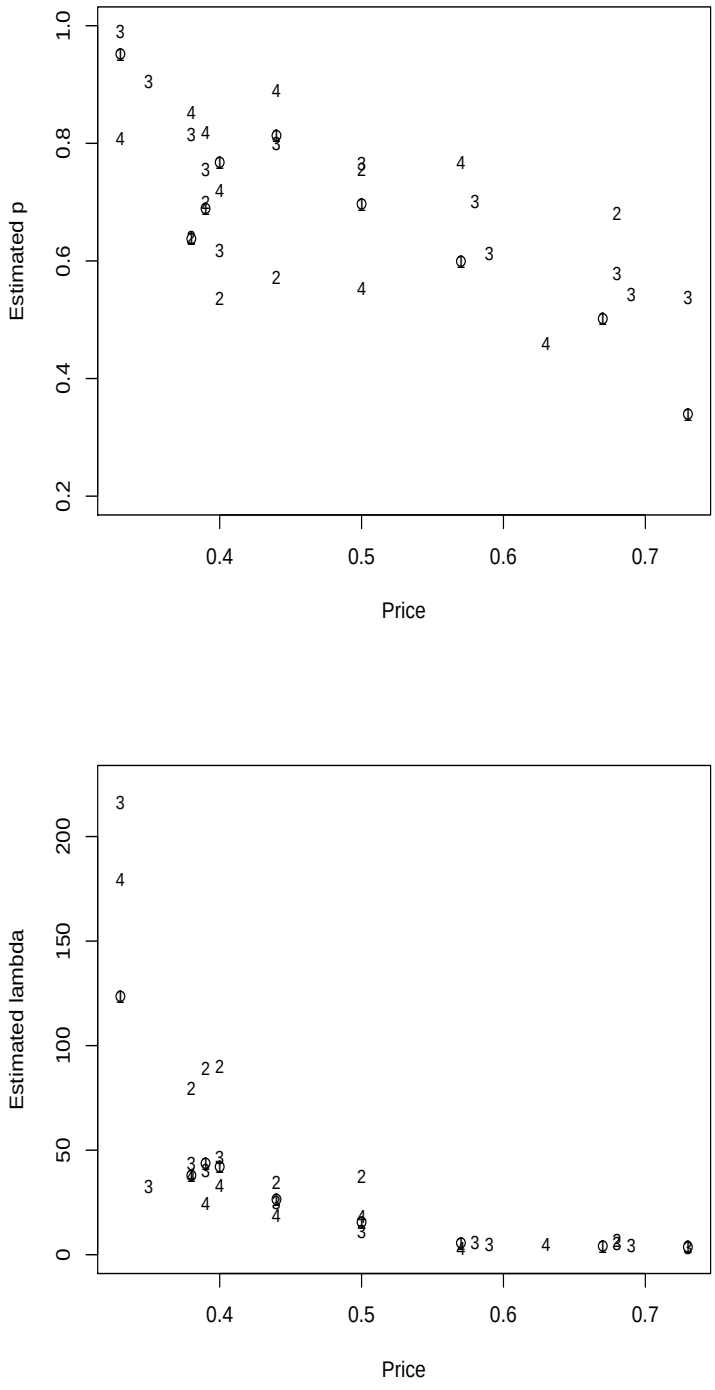


Figure 5: Scatterplots of moment estimates $\hat{p}$ (upper panel) and $\hat{\lambda}$ (lower panel) over price for the four example stores. Values plotted are identification number of store, except that overlapping values for two or more stores are represented as a solid circle.

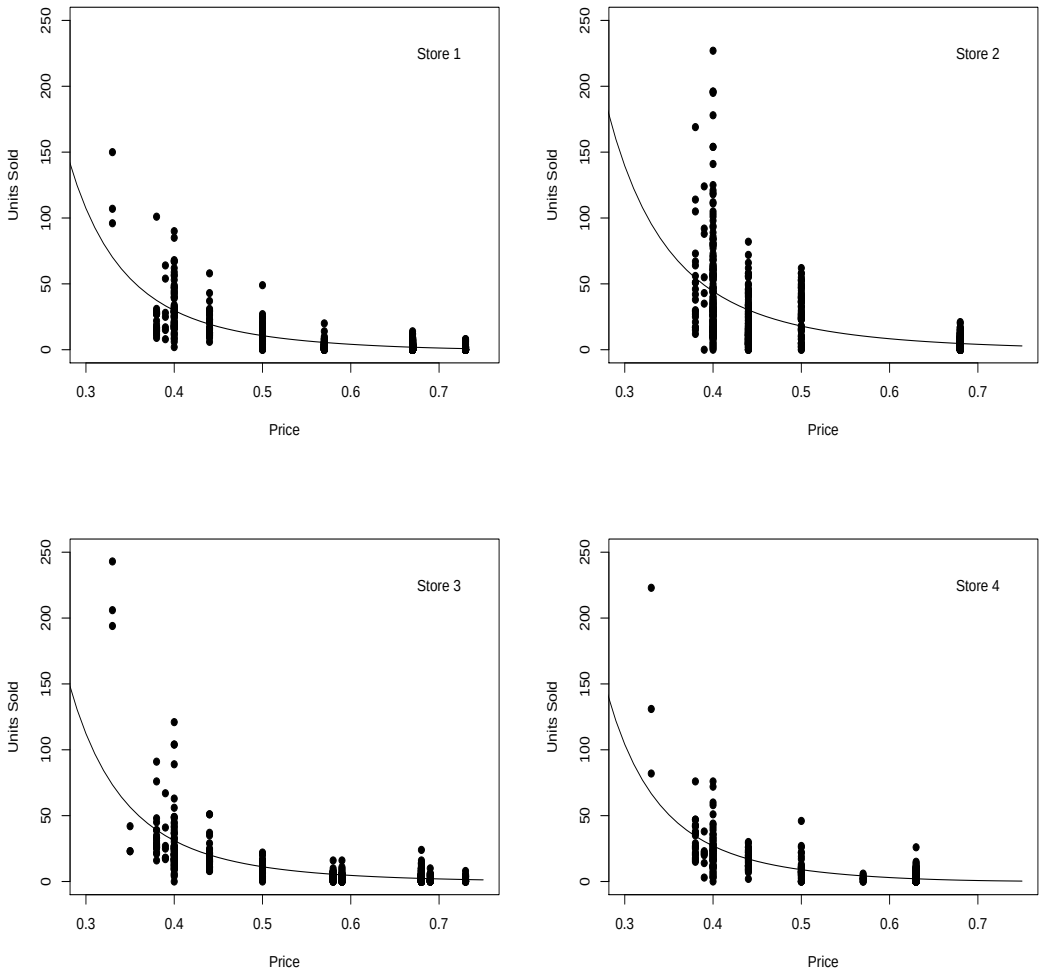


Figure 6: Fitted models from expressions (2) and (4) for the four example stores.

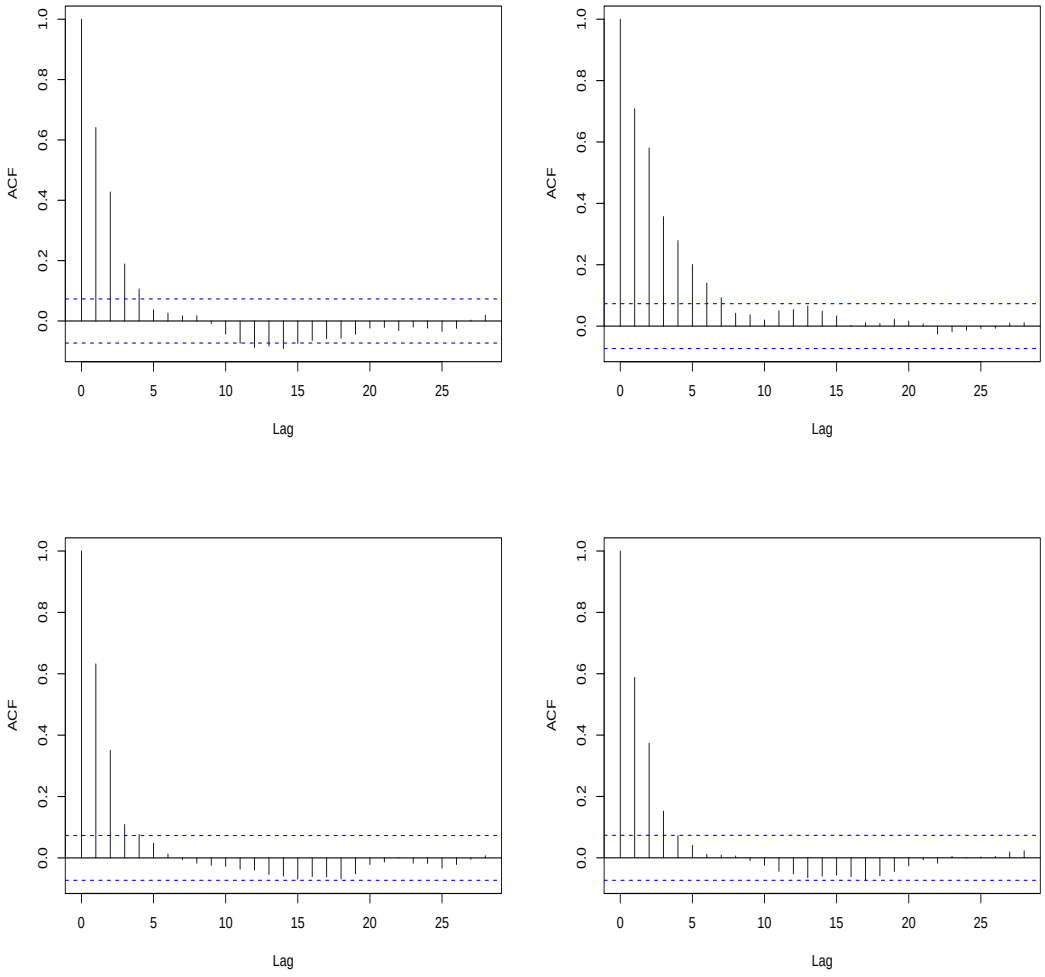


Figure 7: Autocorrelation functions for number of units sold for the four example stores.

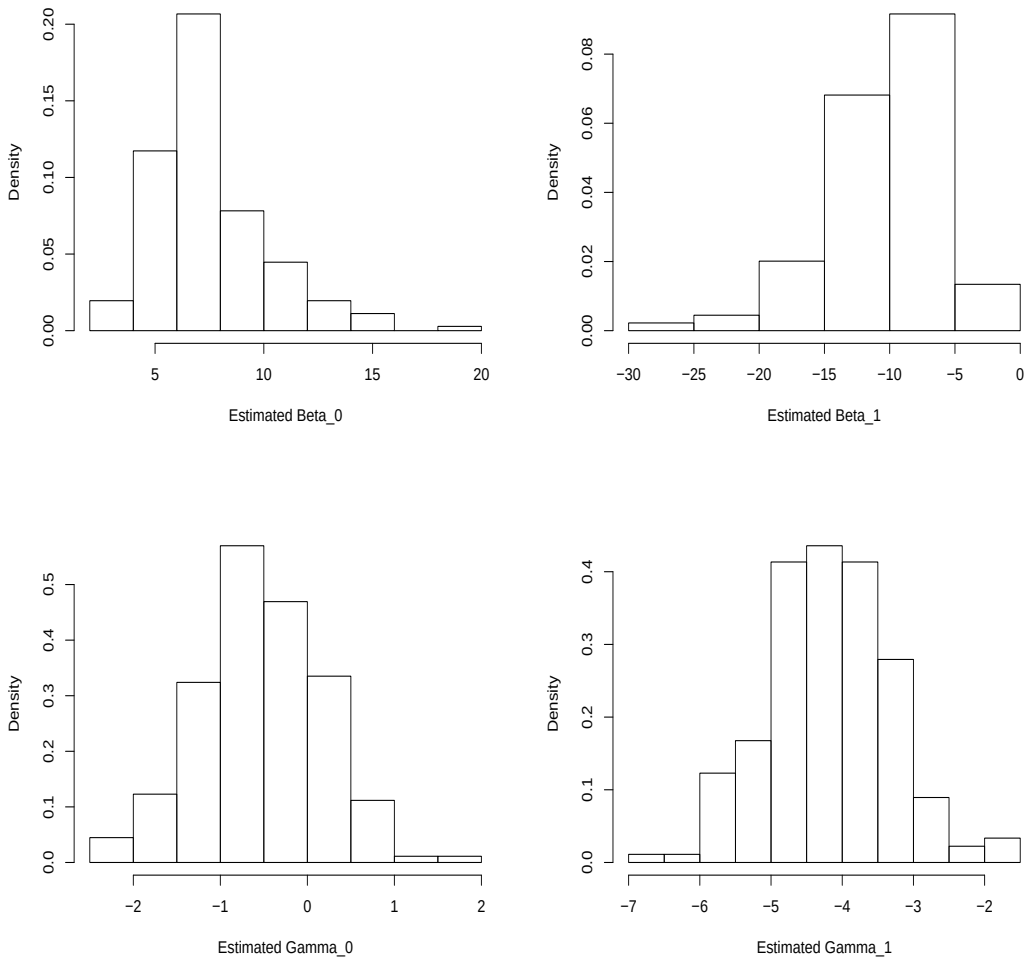


Figure 8: Histograms of estimated values of β_0 (upper left), β_1 (upper right), γ_0 (lower left) and γ_1 (lower right) from separate fits of the basic model to data from individual stores.

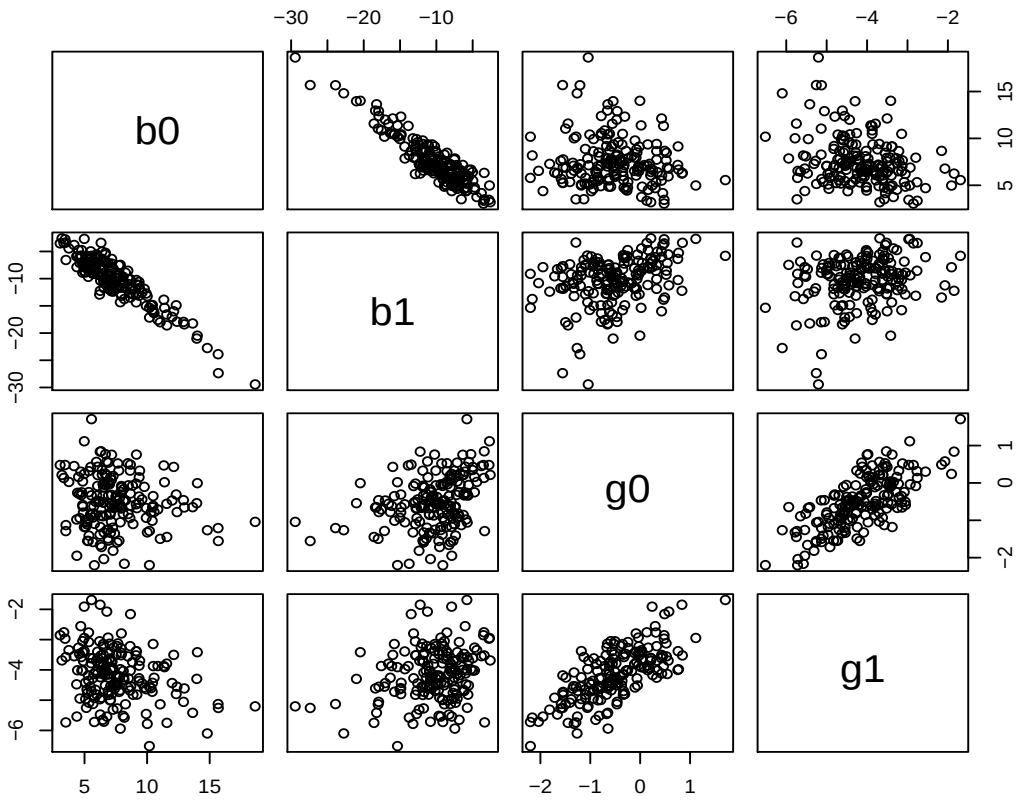


Figure 9: Scatterplot matrix of estimated parameters from separate fits to data from individual stores. In the figure b0 represents $\hat{\beta}_0$, b1 represents $\hat{\beta}_1$, g0 represents $\hat{\gamma}_0$ and g1 represents $\hat{\gamma}_1$.

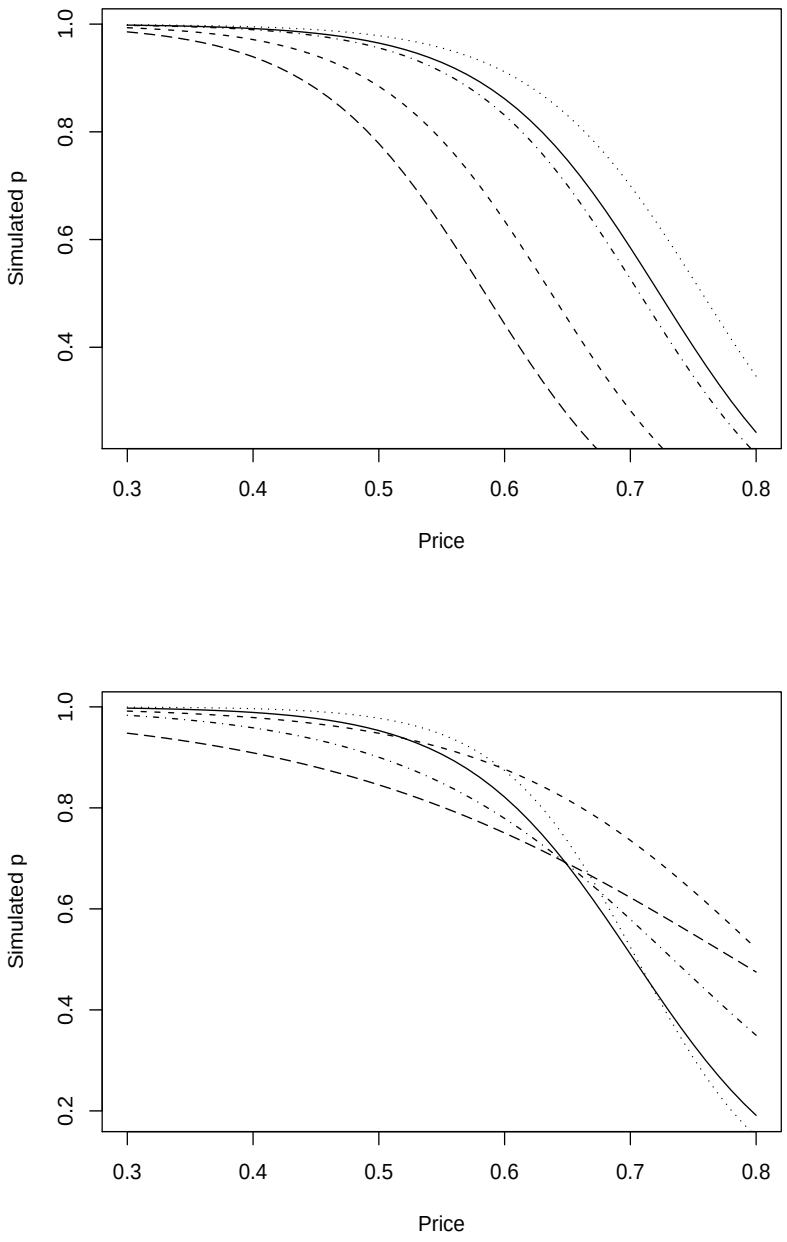


Figure 10: Simulated relations between $p_{i,j}$ and $x_{i,j}$ for Model 7 (upper panel) and Model 8 (lower panel). Each curve represents a randomly selected regression from the appropriate model.

Questions

1. Consider sales of green beans on one day in one store. As in the background material, connect Y with the number of units sold. Also define a latent binary random variable Z to represent the construct of what might be called "consumer interest", which is necessary, but not sufficient, for sales to occur. Using Z and Y, give a mathematical argument that leads to expression (1) as a model for Y.

2. Find the expected value and variance of Y. How does the variance of this distribution compare with that of a "corresponding" Poisson distribution (i.e., a Poisson distribution with the same expected value)?

3. Taking a set of random variables $\{Y_i : i = 1, \ldots, m\}$ as independent and identically distributed with probability mass function (1), find moment estimators of the parameters p and λ.

4. In selecting functional forms for $g_p(\cdot)$ and $g_\lambda(\cdot)$ of (3) we would like functions that allow these parameters to vary with price (x_i) in ways that match the plots of Figure 5. Give one additional consideration that is important in determining specific forms for these functions.

5. Suppose that a Newton-type algorithm was used to approximate maximum likelihood estimates of β_0, β_1, γ_0 and γ_1 in the model of expressions (2) and (4) using data from an individual store. Let $\{Y_i : i = 1, \ldots, n\}$ represent response variables and $\{x_i : i = 1, \ldots, n\}$ represent covariates (price) for this store.

   (a) Write the log likelihood function $\ell(\beta_0, \beta_1, \gamma_0, \gamma_1)$.

   (b) A number of derivatives will be needed for estimation. Write the first partial derivatives of the log likelihood with respect to γ_0 and γ_1.

   (c) If the Newton-type algorithm used proves highly sensitive to starting values (i.e., will not converge for many starting values) what might be done to investigate whether this problem is "caused" by only one (or maybe two) of the parameters in the model?

6. Visually examine the plots of Figure 6 in the question. What aspect of the problem does it appear that the systematic model component is not representing as well as we might like?

7. Raw residuals computed from the model fits of Figure 6 will certainly show unequal variances. The model specifies variances that are unequal depending on price. For basic generalized linear models, or other models based on exponential dispersion family distributions, deviance residuals automatically adjust the variance of residuals for the relation between means and variances dictated by the response distributions. That is not possible here because the model random component (response distributions) cannot be represented as exponential dispersion family distributions. Outline a procedure in terms of steps (i.e., step 1: compute *some quantities* and so on) that one might use to examine the agreement or disagreement of observed variance with model-specified variance.

   Hint: recall that in Question 2 you found the form of the variance for the response distributions on which this model is based. Now, we have obtained maximum likelihood estimates $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$ *and* $\hat{\gamma}_1$ *from fitting the model to data from an individual store. Finally, recall from Figure 1 that replicate values are available at each unique value of price.*

8. Suppose that it is determined that observed variances are greater than dictated by the model, at least for lower prices and larger expected values for number of units sold (this is actually true for these data in most stores). How might you modify the model to account for this phenomenon?

   Hint: recall from Figure 3 that the same price occurs on many different days scattered across the entire time span of the observed data records.

9. The autocorrelation plots of Figure 7 suggest that data on numbers of units sold over time exhibit structure that might be represented though the incorporation of some type of temporal dependence in the model. In a few brief sentences explain why these plots do not indicate that we *must* include temporal dependence in a model in order for it to be an adequate representation of the data structure. Note that this does not imply that we definitely do not need to worry about this aspect of the data structure.

10. Consider the model of expressions (2) and (6). This would not be a good model at all. Why not?

    Note: A one sentence answer is sufficient here.

11. Now consider the model of (2) and (7) versus that of (2) and (8). Our present

concern is only with the $p_{i,j}$. The model of (7) has the form of a generalized linear mixed model, namely,

$$\log\left(\frac{p_{i,j}}{1-p_{i,j}}\right) = \beta_0 + \beta_1 x_{i,j} + \epsilon_j.$$

The model of (8) has the form of a more general random parameter or hierarchical model. Is there any reason to prefer one of the models (7) or (8) over the other?

12. Consider using the model of expression(8). Give a generic construction for the joint distribution of the parameter vector $(\beta_{0,j}, \beta_{1,j}, \gamma_{0,j}, \gamma_{1,j})$. For example, using $g(\cdot)$ as a generic probability density function, $g(x)$ is the density of a random variable X, $g(x, y)$ is a joint density for X and Y, and so forth.

13. Also with respect to the model of (8), identify an issue in the selection of the random parameter distribution $(\beta_{0,j}, \beta_{1,j}, \gamma_{0,j}, \gamma_{1,j})$ that will require either a compromise in model formulation, or an effort to learn about distributions other than those commonly covered in courses such as 500, 511, 542, 543 and 601.

14. It is sometimes claimed that by taking a Bayesian approach to analysis we can circumvent the need to worry about issues such as that you may have identified in the previous question. This claim is based on the following formulation. Consider the overall model as consisting of a data model $f(\boldsymbol{y}|\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1)$, a level 1 prior $\pi_1(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1|\boldsymbol{\theta})$ and a level 2 prior (or hyperprior) $\pi_2(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes any parameters that are contained in the level 1 prior π_1. We are free to choose both the level 1 and level 2 prior distributions as product forms. In particular, using $\pi(\cdot)$ as a generic prior distribution, we could choose the level 1 prior as (now suppressing dependence on $\boldsymbol{\theta}$)

$$\pi_1(\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}_0, \boldsymbol{\gamma}_1) = \pi(\boldsymbol{\beta}_0)\,\pi(\boldsymbol{\beta}_1)\,\pi(\boldsymbol{\gamma}_0)\,\pi(\boldsymbol{\gamma}_1) = \prod_{j=1}^{S} \pi(\beta_{0,j})\,\pi(\beta_{1,j})\,\pi(\gamma_{0,j})\,\pi(\gamma_{1,j}),$$

where the individual univariate distributions are selected based primarily on mathematical convenience.

   (a) Give a brief argument in support of this notion, assuming that the joint posterior distribution will be obtained through simulation using Markov Chain Monte Carlo methods.

   (b) Give a brief argument in contradiction of this notion.

Hint: In part (b) consider whether we might want to predict anything about new stores that might be added to the grocery chain in the future.

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. The probability mass function of expression (1) in the question may be developed as follows. Define a latent random variable Z connected with the construct of "consumer interest", which is necessary, but not sufficient, for sales to occur. That is, let

$$Z_i = \begin{cases} 1 & \text{if there is consumer interest,} \\ 0 & \text{otherwise.} \end{cases}$$

and assign Z a binary probability mass function, for $0 < p < 1$,

$$g(z|p) = \begin{cases} p^z\,(1-p)^{1-z} & z = 0, 1 \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Let Y denote a random variable connected with the number of units sold so that $Y \in \{0, 1, \ldots\}$. Specify a conditional probability mass function for Y as, for $\lambda > 0$,

$$f(y|z,\lambda) = \begin{cases} I(y = 0) & \text{if } z = 0 \\ \frac{1}{y!}\,\lambda^y\,\exp(-\lambda)\,I(y \geq 0) & \text{if } z = 1. \end{cases} \tag{2}$$

From (1) and (2) the joint probability mass function of Z and Y is,

$$m(z,y|p,\lambda) = \begin{cases} 1-p & z = 0,\, y = 0 \\ p\,\frac{1}{y!}\,\lambda^y\,\exp(-\lambda) & z = 1,\, y = 0, 1, \ldots \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Summing (3) over $z = 0, 1$ for each possible value of Y gives expression (1) in the question.

Question 2. The expected value of Y is found as

$$E(Y) = E\{E(Y|Z)\} = 0g(0|p) + \lambda g(1|p) = p\lambda.$$

The variance may be found through a standard conditioning argument but is perhaps most easily derived by first finding $E(Y^2)$ as

$$E(Y^2) = E\{E(Y^2|Z)\} = 0g(0|p) + (\lambda + \lambda^2)g(1|p) = p(\lambda + \lambda^2),$$

which then gives

$$var(Y) = E(Y^2) - \{E(Y)\}^2 = p\lambda(1 + \lambda - p\lambda).$$

The variance of a "corresponding" Poisson distribution would be $\tilde{V} = p\lambda$ and the difference in variances is

$$\begin{aligned} var(Y) - \tilde{V} &= p\lambda(1 + \lambda - p\lambda) - p\lambda \\ &= p\lambda(\lambda - p\lambda) > 0 \end{aligned}$$

since $\lambda > 0$ and $0 < p < 1$.

Question 3. Let $\bar{Y} = (1/m)\sum_{i=1}^m Y_i$ and $S^2 = (1/(m-1))\sum_{i=1}^m (Y_i - \bar{Y})^2$ be the usual sample moments. Moment estimators of p and λ are then given by

$$\begin{aligned} \bar{Y} &= p\lambda \\ S^2 &= p\lambda(1 + \lambda - p\lambda). \end{aligned}$$

Algebra then gives

$$\begin{aligned} \hat{\lambda} &= \frac{S^2}{\bar{Y}} - 1 + \bar{Y} \\ \hat{p} &= \left[\frac{S^2}{\bar{Y}} - \frac{1}{\bar{Y}} + 1\right]^{-1} \end{aligned}$$

Question 4. We would also like functional forms for $g_p(\cdot)$ and $g_\lambda(\cdot)$ that automatically enforce the required parameter spaces $0 < p < 1$ and $\lambda > 0$.

Question 5. (a) Directly from the probability mass function (1) in the question, an individual term of the log likelihood is, after dropping a term that depends only on y_i,

$$\begin{aligned} \ell_i(\beta_0, \beta_1, \gamma_0, \gamma_1) &= \log\left[(1 - p_i) + p_i \exp(-\lambda_i)\right] I(y_i = 0) \\ &+ \log\left[\log(p_i) + y_i \log(\lambda_i) - \lambda_i\right] I(y_i > 0), \end{aligned}$$

where p_i and λ_i are given in expression (4) of the question. By independence the log likelihood is then

$$\ell(\beta_0, \beta_1, \gamma_0, \gamma_1) = \sum_{i=1}^{n} \ell_i(\beta_0, \beta_1, \gamma_0, \gamma_1).$$

(b) The first partial derivatives of a log likelihood term $\ell_i = \ell_i(\beta_0, \beta_1, \gamma_0, \gamma_1)$ with respect to γ_0 and γ_1 are

$$\begin{aligned} \frac{\partial \ell_i}{\partial \gamma_0} &= \frac{\partial \ell_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \gamma_0} \\ \frac{\partial \ell_i}{\partial \gamma_1} &= \frac{\partial \ell_i}{\partial \lambda_i} \frac{\partial \lambda_i}{\partial \gamma_1}. \end{aligned} \tag{4}$$

In (**??**),

$$\begin{aligned} \frac{\partial \ell_i}{\partial \lambda_i} &= \left(\frac{y_i}{\lambda_i} - 1\right) I(y_i > 0) - \frac{p_i \exp(-\lambda_i)}{(1 - p_i) + p_i \exp(-\lambda_i)} I(y_i = 0), \\ \frac{\partial \lambda_i}{\partial \gamma_0} &= \exp\{\gamma_0 + \gamma_1 \log(x_i)\} \\ \frac{\partial \lambda_i}{\partial \gamma_1} &= \exp\{\gamma_0 + \gamma_1 \log(x_i)\} \log(x_i). \end{aligned}$$

The first partial derivatives of the complete log likelihood are then sums of (**??**).

(c) Examine "slices" of the log likelihood function, obtained by holding all but one of the parameters fixed at specified values and plotting the resulting one dimensional function over values of the parameter allowed to vary. Doing this for each parameter in turn will help determine if there is one (or maybe two) dimensions of the likelihood or log likelihood surface that are relatively flat, have multiple local maxima, or otherwise might cause problems for an iterative algorithm that depends on sequences of linear approximations to solve a nonlinear optimization problem.

Question 6. It appears at least visually that the model is having a difficult time picking up the highest expected values (at the lowest price) for at least some of the stores.

Question 7. The model specifies variances as $var(Y_i) = p_i\lambda_i(1 + \lambda_i - p_i\lambda_i)$, and maximum likelihood estimates are available for β_0, β_1, γ_0 and γ_1. A procedure to examine the

agreement or disagreement of observed and model specified variances is then as follows.

1. Let $\{x_k : k = 1, \ldots, K\}$ denote the set of distinct covariate values (price levels) available in the data from a given store. Compute maximum likelihood estimates of p_k and λ_k for $k = 1, \ldots, K$ using the relations given in expression (4) of the question, namely,

$$\begin{aligned} \hat{p_k} &= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_k)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_k)} \\ \hat{\lambda}_k &= \exp[\hat{\gamma}_0 + \hat{\gamma}_k \log(x_k)]. \end{aligned}$$

2. Compute estimates of the variances for $k = 1, \ldots, K$ as

$$\hat{v}_k = \hat{p}_k \hat{\lambda}_k (1 + \hat{\lambda}_k - \hat{p}_k \hat{\lambda}_k).$$

3. Compute sample means and variances for each price level as

$$\begin{aligned} \bar{y}_k &= \sum_{i:x_i = x_k} y_i \\ s_k^2 &= \frac{1}{m_k - 1} \sum_{i:x_i = x_k} (y_i - \bar{y}_k)^2, \end{aligned}$$

where m_k is the number of observations at price level x_k.

4. Plot the $\{\hat{v}_k : k = 1, \ldots, K\}$ against the $\{s_k^2 : k = 1, \ldots, K\}$ or any monotone transformation applied to both sets of values (e.g. logarithms). Examine this plot for equal values.

Question 8. Conditional on consumer interest ($z_i = 1$) the model takes the values of Y_i that share a given covariate value to be independent and identically distributed according to a Poisson distribution. The values of Y_i that share a covariate value can occur on many different days and this motivates the use of different Poisson distributions even for those values of Y_i. One way to accomplish this in the model is to replace the Poisson probability mass function that appears in (2) with a Gamma-Poisson mixture distribution. The intuition that accompanies this suggestion is that, conditional on

consumer interest and at a given price, the number of units sold will have a Poisson distribution randomly selected from a set of Poisson distributions for days with the given price. The marginal probability mass function of Y_i would then become

$$f_i(y|p_i,\alpha_i,\beta_i) = \begin{cases} (1-p_i) + p_i \dfrac{\beta_i^{\alpha_i}}{(\beta_i+1)^{\alpha}} & y = 0 \\ p_i \dfrac{\beta_i^{\alpha_i}\Gamma(\alpha_i+y)}{(\beta_i+1)^{\alpha_i}\Gamma(\alpha_i)y!} & y = 1,2,\ldots \\ 0 & \text{otherwise} \end{cases}$$

To complete the model we would then need to take either $\alpha_i = \alpha$ and $\mu_i = \alpha/\eta_i$ (more likely) or $\beta_i = \beta$ and $\mu_i = \alpha_i/\beta$ (less likely) and $\mu_i = h(x_i, \boldsymbol{\gamma})$ for some known function $h(\cdot)$ and unknown parameter $\boldsymbol{\gamma}$.

Note: the form of the probability mass function for the Y_i was not expected in this answer, just the idea of a Gamma-Poisson mixture and the motivation that additional variability might come from the different days on which a given price occurs.

Question 9. That the response quantities exhibit serial correlation does not imply that we *must* incorporate some type of temporal dependence structure in the model. The covariates of price also exhibit temporal structure, as evidenced by Figure 3 in the question. The effect of the covariates, then, may produce temporal structure in the response variables. We would be better served by examining residual quantities for remaining temporal structure.

Question 10. Because the additive error terms, having possible values of any real number, do not enforce the necessary parameter spaces on the p_i and λ_i.

Question 11. Based on collections of parameter estimates from fitting stores individually, the model of (8) would seem to be preferable to that of (7). That model (7) can be written in the form of a generalized linear mixed model indicates that it would be similar to model (8) if $\beta_{1,j}$ were taken as fixed in that model (as β_1) and only $\beta_{0,j}$ were allowed to be random. This does not seem to be in concert with the histograms of Figure 8 in which both β_0 and β_1 display considerable variability across stores, nor with the patterns of Figure 10 in which curves for $p_{i,j}$ vary in shape for the actual data, but are given as translations by model (7).

Question 12. The scatterplot matrix of Figure 9 suggests that $\beta_{0,j}$ and $\beta_{1,j}$ should receive a joint model, and similarly for $\gamma_{0,j}$ and $\gamma_{1,j}$. There does not appear, however, to be strong (at least linear) relations for parameters between these pairs. Thus, a generic form of random parameter model would be to take

$$g(\beta_{0,j}, \beta_{1,j}, \gamma_{0,j}, \gamma_{1,j}) = g(\beta_{0,j}, \beta_{1,j}|\boldsymbol{\theta}_1)\, g(\gamma_{0,j}, \gamma_{1,j}|\theta_2)$$

Question 13. If $\beta_{0,j}$ and $\beta_{1,j}$ are jointly modeled with a bivariate distribution, it may be difficult to also achieve a marginal distribution for $\beta_{0,j}$ that is skew right and a marginal distribution for $\beta_{1,j}$ that is skew left. We may need to determine which of these aspects of the data we believe is more important to performance of the model.

Note: An alternative, that was not expected as part of the answer to this question, would be to investigate the used of what are sometimes called "skew-normal" or "skew-elliptical" distributions (e.g., search for papers by Azzalini on skew distributions). There are some questions that might arise, depending on the approach taken to estimation (conditions under which likelihood estimates are not finite, see a book edited by Marc Genton, "Skew-Elliptical Distributions and Their Applications").

Question 14. (a) Specification of a joint prior in product form does not imply that posterior distributions for the individual parameter components $\beta_{0,j}$, $\beta_{1,j}$, $\gamma_{0,j}$ and $\gamma_{1,j}$ will be independent within levels of $j = 1, \ldots, S$. Any dependencies that may exist in the posteriors may be examined by looking at sample correlations computed for the simulated values over MCMC iterations.

(b) If we wish to make predictions or forecasts for future stores we would be likely to do so based on the posterior predictive distributions of $\beta_{0,j}$, $\beta_{1,j}$, $\gamma_{0,j}$ and $\gamma_{1,j}$. This will involve simulation of values from the distributions of these parameters as specified in the model. If those distributions do not incorporate correlation between, for example, the $\beta_{0,j}$ and $\beta_{1,j}$, then the resulting forecasts may be misleading.

A certain chemical reaction process is used to generate "product," and the amount of that product resulting from one process run is a function of the time the process is allowed to continue and the temperature at which it is carried out. It is important for the engineers operating the process to understand the relationship among these variables; while it is in their interest to run the process under conditions that yield more (rather than less) product in each run, they also want to limit reaction time (to keep run-times short) and temperature (to limit energy consumption). In order to estimate the functional relationship between these variables, they performed an experiment in which time and temperature were set at various reasonable levels, and the resulting amount of product was measured (with some random errors, assumed independent of all others in each case). The selected combinations of time and temperature, their standardized (or "coded") versions x_1 and x_2, and the resulting measured responses are given in the following table. Note that three replicate runs were made for each of the first four settings of independent variables listed in the table; a total of $n = 17$ independent response realizations were collected as data in the experiment.

Natural Variables		Coded Variables		
Time(min.)	Temperature(°C)	x_1	x_2	Response y (100 lbs.)
80	170	-1.0	-1.0	$n = 3$, $\bar{y} = 2.6182$, $S^2 = 0.029351$
80	180	-1.0	1.0	$n = 3$, $\bar{y} = 5.4231$, $S^2 = 0.025151$
90	170	1.0	-1.0	$n = 3$, $\bar{y} = 4.5884$, $S^2 = 0.015069$
90	180	1.0	1.0	$n = 3$, $\bar{y} = 6.8654$, $S^2 = 0.008321$
92	175	1.4	0.0	6.151196
78	175	-1.4	0.0	3.912642
85	182	0.0	1.4	6.619413
85	168	0.0	-1.4	3.160126
85	175	0.0	0.0	4.857220

(Note that, over all runs in the experiment, the totals of x_1 values, x_2 values, and values of the product x_1x_2, are each zero.) In the analysis following the experiments, the engineers fit three different models to their data:

"Model L":	$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
"Model Q":	$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$
"Model L^p":	$y = (\mu + \beta_1 x_1 + \beta_2 x_2)^p + \epsilon$

Some of the results from these least-squares fits are as follows:

Sample variance of all data values = 2.312523

	Sum of Squared	Parameter Estimates						
Model	Residuals	$\hat{\mu}$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_{12}$	$\hat{\beta}_{11}$	$\hat{\beta}_{22}$	$\hat{p}$
L	0.421257	4.8933	0.8399	1.2618	-	-	-	-
Q	0.191016	4.8625	0.8399	1.2618	-0.1320	0.0527	-0.0198	-
L^p	0.207123	12.39484	3.11428	4.71621	-	-	-	0.64083

1. The engineers wish to test, at level $\alpha = 0.05$, whether second order terms are actually needed in a linear model, i.e.:

$$H_0 : \text{Model } L$$
$$H_A : \text{Model } Q$$

   Compute a test statistic, and completely identify the appropriate critical value or values.

2. The engineers wish to test, at level $\alpha = 0.05$, whether the quadratic polynomial model is actually adequate, i.e.:

$$H_0 : \text{Model } Q$$
$$H_A : E(y) \text{ is an unspecified function of } x_1 \text{ and } x_2$$

   Compute a test statistic, and completely identify the appropriate critical value or values.

3. The engineers are willing to assume that y is normally distributed for any value of (x_1, x_2). They are also confident that $Var(y)$ is not influenced by x_2, but wish to test, at level $\alpha = 0.05$, whether it is also unaffected by the value of x_1, i.e.:

$$H_0 : Var(y) \text{ is constant for all } x_1 \text{ and } x_2$$
$$H_A : Var(y) \text{ is constant for all } x_2$$

   Compute the test statistic, assuming no particular functional form for $E(y)$ as a function of either independent variable, and completely identify the appropriate critical value or values.

4. Under the assumptions that Model Q is adequate, and that $Var(y)$ is constant, compute:

   (a) an estimate of the difference between expected response at $(x_1, x_2) = (1, 1)$ and $(x_1, x_2) = (-1, -1)$, i.e., $E[y(1, 1)] - E[y(-1, -1)]$.

   (b) the standard error of the estimate from (a).

5. Suppose (only for this part) that the data are actually generated by a process accurately described by Model Q *plus* one additional polynomial term, $\beta_{111}x_1^3$. In the fit of Model Q:

   (a) which coefficient estimates (if any) are biased, and what is the bias of these estimates?

   (b) what is the bias (if any) of $\hat{y}(1, 1)$ (the fitted value of y at $(x_1, x_2) = (1, 1)$)?

6. The power transformation is a popular data analytic tool in the modeling of positive-valued data; it is often used in an attempt to find a "scale" on which responses appear to have a simple mean structure (as a function of the independent variables) and a constant variance. Our engineers might consider fitting y^q to Model L, for several values of q.

   (a) Is fitting y to Model L^p equivalent to finding the value of $(\alpha, \beta_1, \beta_2, q)$ that minimizes the sum of squared differences between y^q and the mean response function of Model L? Why or why not?

   (b) Is it reasonable to fit y^q to Model L for several values of q, and select the value of q that results in the smallest MSE for a "final" analysis of the data? Why or why not?

   (c) Given what you've been told about the data and least-squares fits of models L and Q, what would be good starting values for a nonlinear least-squares fit to model L^p?

A statistical package was used to fit model L^p to the data, with satisfactory numerical results (i.e. apparent convergence, etc.). Least-squares estimates of the four model parameters are given on page 1, and the estimated variance covariance matrix of these estimators is:

$$\widehat{Var}\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{p} \end{pmatrix} = \begin{pmatrix} 9.9790276 & 3.1990299 & 4.9190147 & -0.1963993 \\ 3.1990299 & 1.0386598 & 1.5757804 & -0.0629945 \\ 4.9190147 & 1.5757804 & 2.4378424 & -0.0968665 \\ -0.1963993 & -0.0629945 & -0.0968665 & 0.0038707 \end{pmatrix}.$$

7. Based on this information, compute the following quantities. (You may find it helpful to recall that $\frac{\partial}{\partial t}a^t = a^t \times ln(a)$. Warning: Be very careful with the arithmetic here.)

   (a) an estimate of the expected response at $(x_1, x_2) = (0, 0)$.

   (b) an approximate standard error of the estimate from (a).

   (c) a prediction of the total response (100 lbs. of product) that would be generated from 3 future (and independent) runs of the process at $(x_1, x_2) = (0, 0)$.

   (d) the standard error of the prediction from (c).

1. $SS(\beta_{12}, \beta_{11}, \beta_{22}|\mu, \beta_1, \beta_2) = 0.421257 - 0.191016 = 0.230241$, and the associated MS is $0.230241/3 = 0.076747$. SS(residual) for Q is 0.191016, and the associated MS is $0.191016/(17-6) = 0.017365$. The test statistic, $F = 0.076747/0.017365 = \underline{4.420}$, is compared to the critical value $\underline{F_{0.95}(3, 11)}$.

2. The "pure error" SS can be computed from the sample variances at the replicated points:

$$SSPE = (3-1) \times (0.02935 + 0.025151 + 0.015069 + 0.008321) = 0.155784$$

and the associated MS is $0.155784/8 = 0.019473$. The "lack of fit" SS is the difference between SS(residual) and $SSPE$, $0.191016 - 0.155784 = 0.035232$, and the associated MS is $0.035232/(11-8) = 0.011744$. The test statistic, $F = 0.011744/0.019473 = \underline{0.6031}$, is compared to the critical value $\underline{F_{0.95}(3, 8)}$.

3. Independent estimates of $Var(y|x_1 = -1)$ and $Var(y|x_1 = 1)$ can be constructed by pooling sample variances at the replicated design points:

$$\hat{Var}(y|x_1 = -1) = (0.029351 + 0.025151)/2 = 0.027251$$
$$\hat{Var}(y|x_1 = +1) = (0.015069 + 0.008321)/2 = 0.011695$$

Their ratio, $F = 0.027251/0.011695 = \underline{2.330}$, can be used as the test statistic, with critical values $\underline{F_{0.025}(4, 4)}$ and $\underline{F_{0.975}(4, 4)}$, since the alternative is two-sided.

4. -

   (a) The quantity of interest is equal to:

$$(\mu + \beta_1 + \beta_2 + \beta_{12} + \beta_{11} + \beta_{22}) - (\mu - \beta_1 - \beta_2 + \beta_{12} + \beta_{11} + \beta_{22}) = 2(\beta_1 + \beta_2)$$

   so it can be estimated as $2(0.8399 + 1.2618) = \underline{4.2154}$.

   (b) Columns of the model matrix associated with β_1 and β_2 are orthogonal to each other and to all other columns. Therefore:

$$se^2(\hat{\beta}_1) = se^2(\hat{\beta}_2) = MSE \times (12 \times 1^2 + 2 \times 1.4^2)^{-1} = 0.001091$$

   Because $\hat{\beta}_1$ and $\hat{\beta}_2$ are independent:

$$se(\hat{y}(1, 1) - \hat{y}(-1, -1)) = 2\sqrt{0.001091 + 0.001091} = \underline{0.0934}$$

5. Let $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ be the matrix expression for model Q over the given design. We now suppose that $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_{111}\beta_{111}$, where

$$\mathbf{X}'_{111} = (-1(3 \text{ reps}), -1(3 \text{ reps}), 1(3 \text{ reps}), 1(3 \text{ reps}), 1.4^3, -1.4^3, 0, 0, 0).$$

   (a) Estimates of the model Q parameters have expectation:

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_{111}\beta_{111}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_{111}\beta_{111}.$$

   But $\mathbf{X}_{111}$ is orthogonal to all columns of $\mathbf{X}$ except for that associated with β_1. As a result, only this estimate is biased:

$$E(\hat{\beta}_1) = \beta_1 + (12 \times 1^4 + 2 \times 1.4^4)/(12 \times 1^2 + 2 \times 1.4^2)\beta_{111} = \beta_1 + \underline{1.2364\beta_{111}}$$

   (b) Since $\hat{y}(1,1) = \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_{12} + \hat{\beta}_{11} + \hat{\beta}_{22}$, the bias of $\hat{y}(1,1)$ is $\underline{(1.2364 + 1)\beta_{111}}$.

6. -

   (a) No. The power transform is equivalent to fitting $y = (\mu + \beta_1 x_1 + \beta_2 x_2 + \epsilon)^{1/q}$, which implies (among other things) that $Var(y)$ is potentially a function of x_1 and x_2. Model L^p raises only the "mean structure" to a power, so $Var(y)$ is constant for any p (so long as $Var(\epsilon)$ is constant for each observation).

   (b) No. By changing q, the units of measurement for the data, and so also for the resulting MSE, are being changed. For example, the MSE for $q = \frac{3}{4}$ is $(100 \text{ lbs})^{3/2}$, which cannot be meaningfully compared to the MSE for $q = \frac{1}{4}$, which is $(100 \text{ lbs})^{1/2}$.

   (c) Because model L fits the data reasonably well (with $MSE = 0.421257/(17-3) = 0.03$ out of total $S^2 = 2.31$) and is a special case of model L^p, $\underline{\hat{\mu}_0 = 4.8933,\ \hat{\beta}_{10} = 0.8399,\ \hat{\beta}_{20} = 1.2618,\ \hat{p}_0 = 1.}$ are reasonable starting values.

7. -

   (a) The expected response at $x_1 = x_2 = 0$ is $\eta = \mu^p$, and can be estimated as $12.3948^{0.6408} = \underline{5.0182}$.

   (b) Differentiation w.r.t. the model parameters and evaluating at their estimated values yields:

$$\begin{array}{ll} \frac{\partial}{\partial\mu}\eta = p\mu^{p-1} & 0.6408 \times 12.3948^{-0.3592} = 0.2594 \\ \frac{\partial}{\partial\beta_1}\eta = 0 & \\ \frac{\partial}{\partial\beta_2}\eta = 0 & \\ \frac{\partial}{\partial p}\eta = \mu^p ln(\mu) & 12.3948^{0.6408} ln(12.3948) = 12.6322 \end{array}$$

   So, an approximate variance for $\hat{\eta}$ is:

$$(0.2594 \ \ 12.6322) \begin{pmatrix} 9.9790276 & -0.1963993 \\ -0.1963993 & 0.0038707 \end{pmatrix} \begin{pmatrix} 0.2594 \\ 12.6322 \end{pmatrix} = 0.0020.$$

   So $se(\hat{\eta}) = \sqrt{0.0020} = \underline{0.045}$.

   (c) $3 \times 5.0182 = \underline{15.0546}$.

   (d) $\sqrt{3 \times MSE + 9 \times se^2(\hat{\eta})} = \sqrt{3 \times 0.0148 + 9 \times 0.0020} = \underline{0.2498}$.