PhD Prelim Exam
METHODS

Summer 2014
(Given on 7/15/14)

In an industrial experiment, the effects of three different training programs on the performance of production floor employees of a large manufacturing company were being compared. The three programs are labeled 1, 2, and 3, with 1 representing a traditional training program, the other two programs being modern methods incorporating video instruction and computer applications, respectively. Three non-overlapping random samples of 30 employees were selected and assigned to each program, and a quantitative variable `PreScore` (a measure of innate and acquired skill of each worker which can range from 0 to 100 points) was measured on each employee before the experiment began. The performance level of each worker was measured as a quantitative variable `Score` (which can range from 0 to 200 points) earned over a period of time after the training programs were instituted. Figure 1 shows the `Score` variable plotted against the `PreScore` variable for workers in each program. The three programs represent *levels* of the factor `Program`, the `Score` variable is the *response*, and `PreScore` is a *covariate*.

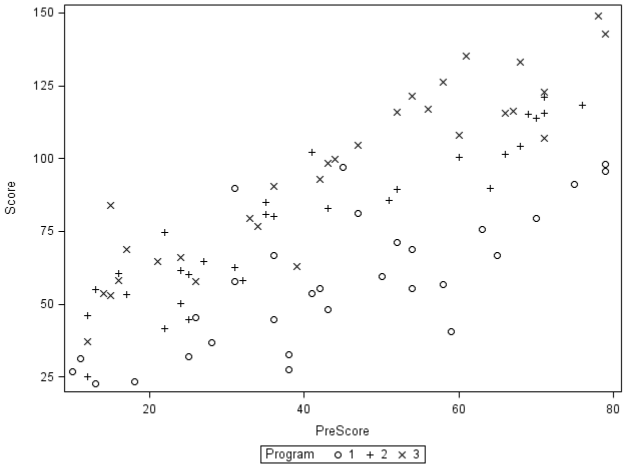


Figure 1: Performance Score plotted against Pre-score of Skill Level

Some summary statistics for these data are given in Table 1.

Program	Variable	N	Mean	Std.Dev.	S.E.(Mean)
1	Score	30	57.693	23.53	4.295
	PreScore	30	43.900	19.47	3.555
2	Score	30	78.123	26.40	4.820
	PreScore	30	40.333	21.22	3.874
3	Score	30	95.250	30.36	5.542
	PreScore	30	43.967	21.31	3.891

Table 1: Some summary statistics

Part A

As an initial analysis, we will examine whether there were any differences among the three training programs without considering the covariate `PreScore`. Side-by-side boxplots of `Score` by `Program` are shown in Figure 2. Let μ_1, μ_2, and μ_3 denote the expected mean scores (population means) from the three programs. We wish to test the hypothesis $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : not H_0.

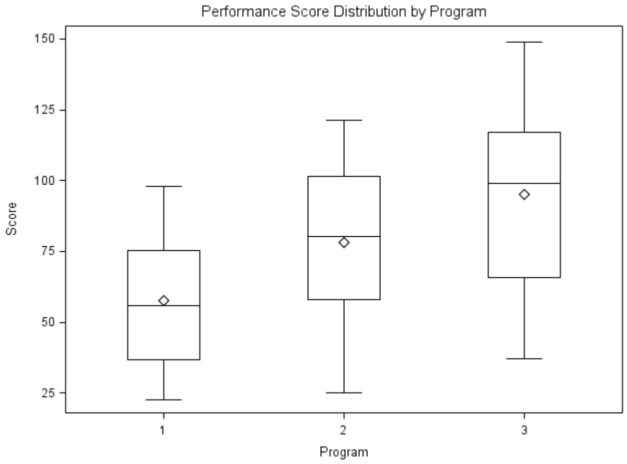


Figure 2: Side-by-side boxplots of Performance by Program (◇ indicate sample means)

1. Using the graph in Figure 2 and the statistics in Table 1, is there evidence to call into question the assumptions necessary to perform a test of the above hypothesis based on an ANOVA table? What other statistical analyses would you use to further verify whether these assumptions are plausible?

2. Complete the following ANOVA table and perform a test of $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : not H_0.

Source	df	SS	MS	F
Between Programs		21212.11		
Within Programs				
Total		84197.64		

3. Calculate a two-sided 90% confidence interval for the linear combination $2\mu_1 - \mu_2 - \mu_3$.

4. Based on the interval in 3., what can be said about a p-value that may be calculated for testing $H_0 : 2\mu_1 - \mu_2 - \mu_3 \geq 0$ vs. $H_a : 2\mu_1 - \mu_2 - \mu_3 < 0$. Explain what this p-value tells the experimenter about the effects of the three programs on the performance levels of the employees.

Part B

A deficiency of the analysis in Part A is that the differences among the 3 groups that may be due to other factors were not controlled. For example, the innate ability of workers being trained may affect their performance, irrespective of the training program. In an analysis of variance setting, including a *covariate* such as `PreScore` may improve the power of the test for detecting differences among the program means.

Consider the following model:

$$y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}, i = 1, 2, 3; j = 1, \ldots, 30 \tag{1}$$

where y_{ij} and x_{ij} represent the `Score` and the `PreScore` obtained by the j^{th} worker in `Program` i, $\epsilon_{ij} \sim iid\ N(0, \sigma^2)$, α_i are the intercepts, and β, a common slope. A fit of this model is illustrated in Figure 3 that shows the lines with the estimated common slope and different intercepts for each program. Use the edited SAS output from a Proc GLM fit of the above model provided in Figure 5 on Page 5 to answer the following questions.

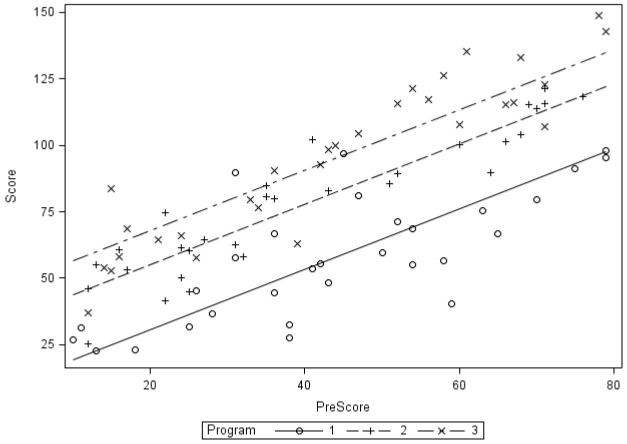


Figure 3: Regression of Performance on PreScore by Program

5. Test a hypothesis to answer the question "Does the covariate `PreScore` contribute to the variabilty in the performance score from the training programs"?

6. Under model (1), the expected mean response for `Program` i at any $x = x^*$ is given by $\mu_i = E(y_{ij}) = \alpha_i + \beta x^*$. Calculate an F statistic to test $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : not H_0. Specify the degrees of freedom for this F test. (Note that this is equivalent to testing that the intercepts of the population regression lines are equal, i.e.,$H_0 : \alpha_1 = \alpha_2 = \alpha_3$ vs. H_a : not H_0.)

7. An estimate of μ_i of 6. at any $x = x^*$ is $\hat{\alpha}_i + \hat{\beta}x^*$ where $\hat{\beta}$ is the least squares estimate of the common slope β (These are called the *adjusted means* and is usually denoted by $\hat{\mu}_i(adj.)$ or $\bar{y}_i(adj)$). Calculate the adjusted means for the three programs at $x = \bar{x}_{..} = 42.7333$.

8. Under model (1), the difference in a pair of expected means of `Score` *adjusted for PreScore* is estimated simply by the corresponding difference in the intercepts (α_i's). Calculate a two-sided 95% confidence interval for $\mu_1 - \mu_3$ using information in the SAS output in Figure 5. Interpret this confidence interval in the context of this experiment.

Part C

In this study, it is possible that how the prior skills of workers affect performance varies from program to program. This possibility is manifested in Figure 4 as unequal slopes in the regression fits. This situation is represented by the following model:

$$y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}, i = 1, 2, 3; j = 1, \ldots, 30 \tag{2}$$

where β_i is the slope for `Program` i, and other quantities remain the same as for model (1). For the purpose of testing the hypothesis that the slopes are in fact different, it is convenient to fit models (1) and (2) using a multiple regression set-up. To fit model (1), indicator variables V1 and V2 are created to represent the `Program` variable with V1=0 and V2=0 when Program=1, V1=1 and V2=0 when Program=2, and V1=0 and V2=1 when Program=3. Fitting the regression model containing the regressors V1, V2, PreScore (and an intercept) is equivalent to fitting model (1). To fit model (2), two new variables INT1 and INT2 representing the products of PreScore with each of V1 and V2, respectively, are created. Fitting the regression model containing the 5 regressors V1, V2, PreScore, INT1, and INT2, (plus an intercept) is equivalent to fitting model (2). The two regression models described above are fitted with `Score` as the response using SAS. The SAS Outputs II and III in Figures 6 and 7, respectively, display the results of these fits.

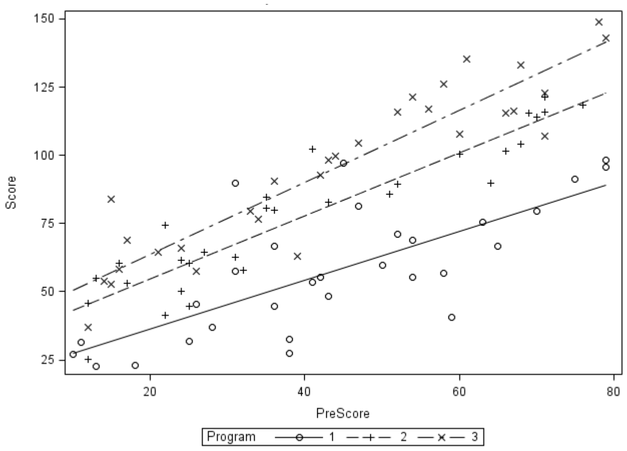


Figure 4: Covariate by Treatment Interaction

9. Use the results in SAS Outputs II and III to construct an F statistic to test the hypothesis that the slopes of the three lines are not all the same. What is your conclusion?

10. The adjusted means when the slopes are unequal can be easily computed as predicted values from the regression fit of model (2) above. Calculate the adjusted means for Programs 1 and 2 when the variable `Prescore` has the value 50.

11. Under model (2), the expected mean response for `Program` i at any $x = x^*$ is given by $\mu_i = E(y_{ij}) = \alpha_i + \beta_i x^*$. Using the regression fit of model (2), calculate a 95% confidence interval for $\mu_1 - \mu_2$ when `Prescore` has the value 50.

SAS Output I

The output below is part of the output produced from executing the following SAS Proc GLM step:

```
proc glm data=training2;
class Program;
model Score=Program PreScore/solution;
run;
```

R-Square	Coeff Var	Root MSE	Score Mean
0.821350	17.17061	13.22518	77.02222

Source	DF	Type I SS	Mean Square	F Value	Pr > F
Program	2	21212.10822	10606.05411	60.64	<.0001
PreScore	1	47943.65496	47943.65496	274.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
Program	2	21726.92466	10863.46233	62.11	<.0001
PreScore	1	47943.65496	47943.65496	274.11	<.0001

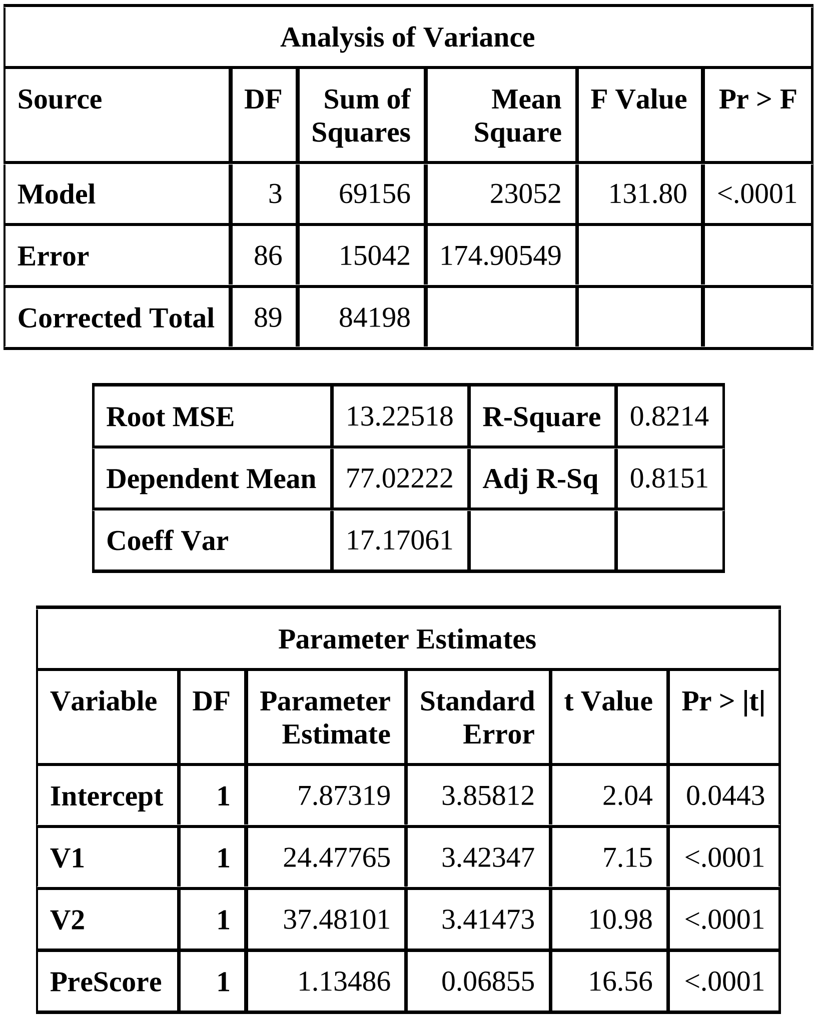
Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	45.35420123	B	3.86168042	11.74	<.0001
Program 1	-37.48100966	B	3.41473089	-10.98	<.0001
Program 2	-13.00335957	B	3.42379771	-3.80	0.0003
Program 3	0.00000000	B	.	.	.
PreScore	1.13485517		0.06854513	16.56	<.0001

Figure 5: Partial Output from Proc GLM

SAS Output II

The output below is part of the output produced from executing the following SAS Proc REG step:

```
proc reg data=training2;
model Score=V1 V2 PreScore;
run;
```



Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	69156	23052	131.80	<.0001
Error	86	15042	174.90549		
Corrected Total	89	84198			

Root MSE	13.22518	**R-Square**	0.8214
Dependent Mean	77.02222	**Adj R-Sq**	0.8151
Coeff Var	17.17061		

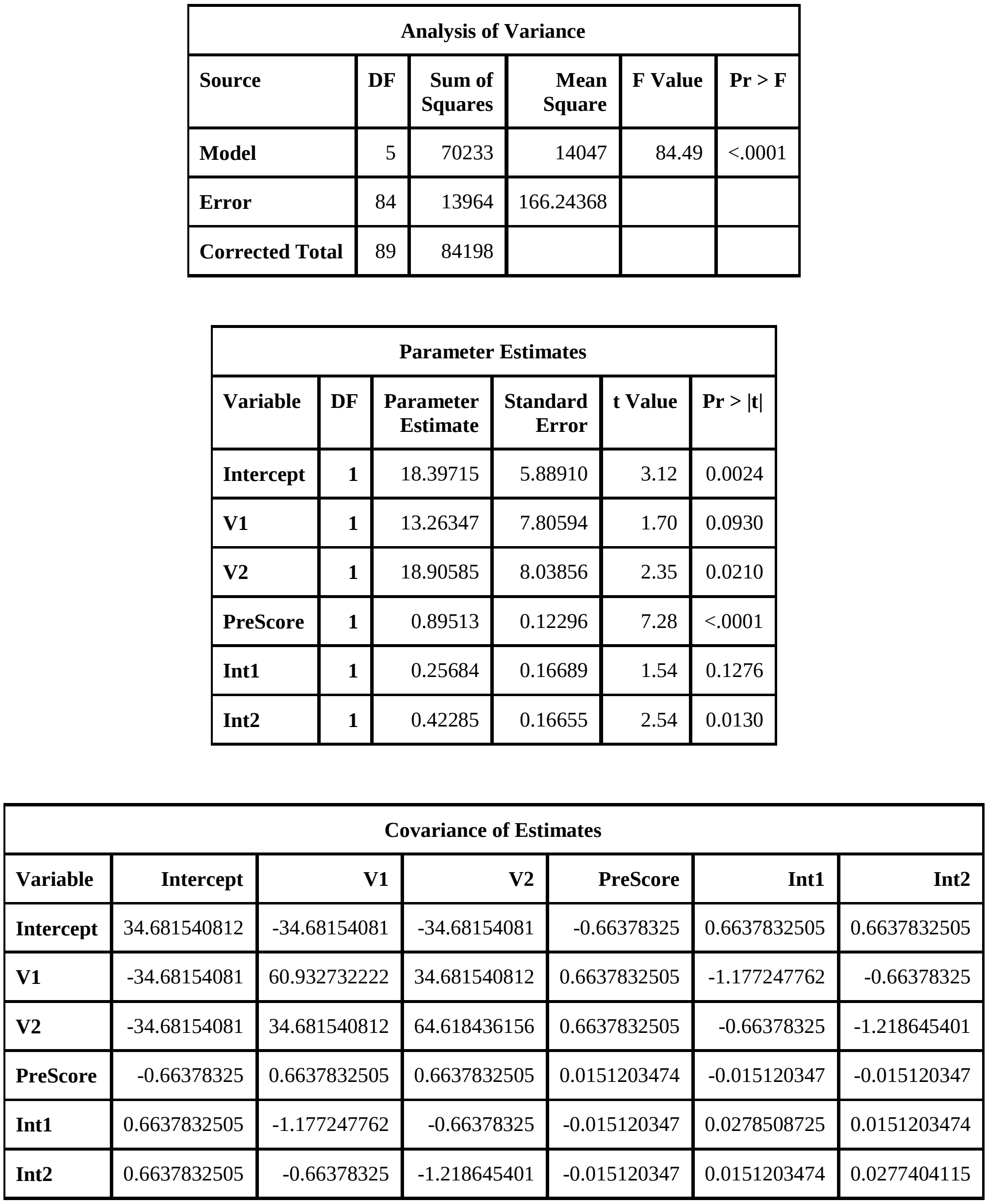
Parameter Estimates					
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**
Intercept	**1**	7.87319	3.85812	2.04	0.0443
V1	**1**	24.47765	3.42347	7.15	<.0001
V2	**1**	37.48101	3.41473	10.98	<.0001
PreScore	**1**	1.13486	0.06855	16.56	<.0001

Figure 6: Output from Proc Reg fit of Model 1

SAS Output III

The output below is part of the output produced from executing the following SAS Proc REG step:

```
proc reg data=training2;
model Score=V1 V2 PreScore INT1 INT2;
run;
```



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	70233	14047	84.49	<.0001
Error	84	13964	166.24368		
Corrected Total	89	84198			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	**1**	18.39715	5.88910	3.12	0.0024
V1	**1**	13.26347	7.80594	1.70	0.0930
V2	**1**	18.90585	8.03856	2.35	0.0210
PreScore	**1**	0.89513	0.12296	7.28	<.0001
Int1	**1**	0.25684	0.16689	1.54	0.1276
Int2	**1**	0.42285	0.16655	2.54	0.0130

Covariance of Estimates

Variable	Intercept	V1	V2	PreScore	Int1	Int2
Intercept	34.681540812	-34.68154081	-34.68154081	-0.66378325	0.6637832505	0.6637832505
V1	-34.68154081	60.932732222	34.681540812	0.6637832505	-1.177247762	-0.66378325
V2	-34.68154081	34.681540812	64.618436156	0.6637832505	-0.66378325	-1.218645401
PreScore	-0.66378325	0.6637832505	0.6637832505	0.0151203474	-0.015120347	-0.015120347
Int1	0.6637832505	-1.177247762	-0.66378325	-0.015120347	0.0278508725	0.0151203474
Int2	0.6637832505	-0.66378325	-1.218645401	-0.015120347	0.0151203474	0.0277404115

Figure 7: Output from Proc Reg fit of Model 2

Part A

1. The side-by-side boxplots of the data show similar spread among programs and little evidence of skewness. There is no strong evidence against normality here. Furthermore the sample standard deviations are close (see the Table 1)). We may formally test the equality of variances across the three groups, say using the `hovtest` option in SAS's *proc anova* (Levene's or Brown-Forsythe's tests) and also test for normality with Shapiro-Wilk or Andesron-Darling statistics, say, using *proc univariate*.

2. The complete Anova table is:

Source	df	SS	MS	F
Between Programs	2	21212.11	10606.05	14.65
Within Programs	87	62985.53	723.97	
Total	89	84197.64		

   Test of $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : at least one pair $\mu_i \neq \mu_j$

   $F_{.05,2,87} \approx 3.11$ Since $F_c = 14.65$ reject H_0 and conclude that there are differences among the expected means of program scores.

3. Under assumption of equal variances, the pooled variance esitimate is $\hat{\sigma}^2 = 723.97$
   For a 90% confidence interval (CI) need $t_{.05,87} \approx 1.66$

   Estimate of $2\mu_1 - \mu_2 - \mu_3$ is: $2 \times 57.693 - 78.123 - 95.250 = -57.987$

   Standard srror of this estimate is: $\sqrt{723.97} \times \sqrt{(2^2 + 1^2 + 1^2)/30} = 12.033$

   Thus a 90% CI for $2\mu_1 - \mu_2 - \mu_3$ is $-57.987 \pm (1.66 \times 12.033 = (-77.96, -38.01)$

4. Test $H_0 : 2\mu_1 - \mu_2 - \mu_3 \geq 0$ vs $H_a : 2\mu_1 - \mu_2 - \mu_3 < 0$ using the 90% CI in Problem#3

   The CI $(-77.96, -38.01)$ is entirely less than zero shows that the p-values of this test is less than .05 Thus we reject H_0 at $\alpha = .05$ and conclude that Programs 2 and 3 do indeed produce a higher average mean score than does Program 1.

Part B

5. Test the hypothesis $H_0 : \beta = 0$ vs. $H_a : \beta \neq 0$ in model (1). Note that β is the coefficient of the control variable `PreScore` in model (1). Using the SAS Output I, the sequential (or SAS Type I) and the partial (or SAS Type II) sums of squares for `PreScore` are the same and provide the correct F statistic for testing this hypothesis. The associated p-value is $< .0001$. Equivalently, we could us the t-test for this coefficient (see lower in the SAS output) to test the above hypothesis. We see the p-value is the same $< .0001$. Thus the conclusion is that *PreScore* does contribute significantly to the variability in `Score`.

6. Test $H_0 : \mu_1 = \mu_2 = \mu_3$ vs. H_a : at least one pair $\mu_i \neq \mu_j$ where μ_i's are defined in model (1) when β is non zero. That is, μ_i's here will be expected mean scores under this model. This is equivalent to testing whether the intercepts are all the same in the model $y_{ij} = \alpha_i + \beta x_{ij} + \epsilon_{ij}$

   The sum of squares for Program adjusted for PreScore (Type III SS) from SAS Output I

   $= 21726.92466$ with 2 df

The Error sum of squares for testing this hypothesis =

Total SS (from Part A) - (Program SS (Type I SS)+ PreScore SS (Type I SS))

=15041.88 with 86 df

That is, we obtain the following partitioning of the Total SS using the Type I SS:

Source	df	SS
Programs	2	21212.11
PreScore	1	47943.65
Error	86	15041.88
Total	89	84197.64

Thus the required F-statistic = (21726.92466/2)/(15041.88/86)=62.11 with 2 and 86 df.

Note that this agrees with the F-statistic given in Type III SS portion of the SAS output where the Program means are *adjusted* for `PreScore`.

7. Use SAS Output I to calculate adjusted means at $\bar{x}_{..}$ using $\bar{y}_i(adj) = \hat{\alpha}_i + \hat{\beta}\bar{x}_{..}$ where α_i is the intercept of the i^{th} line. Using the *parameter estimates* output from GLM, we have

$$\hat{\alpha}_1 = 45.3542 - 37.481 = 7.8732 \text{ and thus } \bar{y}_1(adj) = 7.8732 + 1.13486 \times 42.7333 = 56.369$$

$$\hat{\alpha}_2 = 45.3542 - 13.00336 = 32.3508 \text{ and thus } \bar{y}_2(adj) = 32.3508 + 1.13486 \times 42.7333 = 80.847$$

$$\hat{\alpha}_3 = 45.3542 - 0 = 45.3542 \text{ and thus } \bar{y}_3(adj) = 45.3542 + 1.13486 \times 42.7333 = 93.8505$$

8. We need a 95% confidence interval for $\mu_1 - \mu_3$. Here we can use the *parameter estimates* and their standard errors to calculate this. Clearly $\mu_1 - \mu_3 = \alpha_1 - \alpha_3$ is estimable and the estimate and standard errors are (from SAS Output 1):

$$\widehat{\mu_1 - \mu_3} = \hat{\alpha}_1 - \hat{\alpha}_3 = -37.481 - 0 = -37.481$$

and

$$\text{Standard Error}(\widehat{\mu_1 - \mu_2}) = \text{Standard Error}(\hat{\alpha}_1) = 3.4147$$

Thus the required confidence interval is $-37.481 \pm t_{.025,86} \times 3.4147$ where $t_{.025,86} \approx 1.988$ i.e $(-44.31, -30.6516)$

Part C

9. The hypothesis that the slopes are not equal i.e., $H_0 : \beta_1 = \beta_2 = \beta_3$ vs. H_a : at least one pair $\beta_i \neq \beta_j$ in the model $y_{ij} = \alpha_i + \beta_i x_{ij} + \epsilon_{ij}$ is equivalent to testing whether the variables INT1 and INT2 are significant in the regression fit of Model (1). This is easily accomplished by comparing regression fits of Model (1) and Model (2) using the model comparison F-test:

$$F = \frac{(Error\ SS(Model\ (2)) - Error\ SS(Model\ (1)))/(Error\ DF(Model\ (2)) - Error\ DF(Model\ (1)))}{Error\ SS(Model\ (2))/(Error\ DF(Model\ (1))}$$

$$= \frac{(15042 - 13964)/(86 - 84)}{13964/84} = 539.0/166.24 = 3.24$$

$F_{.05,2,84} \approx 3.11$ Thus we reject the hypothesis of equal slopes at $\alpha = .05$

10. The required adjusted means are easily computed using the regression fit of Model (1) (denoting estimates of the corresponding coefficients for this model by $\hat{\beta}$'s below):

For Program 1: $\hat{\mu}_1(adj) = \hat{y}_1 = \hat{\beta}_0 + 50\hat{\beta}_3 = 18.39715 + 50 \times 0.89513 = 63.15362$

For Program 2: $\hat{\mu}_2(adj) =$
$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 + 50\hat{\beta}_3 + 50\hat{\beta}_4 = 18.39715 + 13.26341 + 50 \times 0.89513 + 50 \times 0.25684 = 89.25902$

11. To compute the required confidence interval, you need an estimate of $\mu_2 - \mu_1$ and an estimate of the variance of this estimate from the fit of the Model (2):

We have $\widehat{\mu_1 - \mu_2} = \hat{\mu}_1(adj) - \hat{\mu}_2(adj) = \hat{y}_1 - \hat{y}_2 = -\hat{\beta}_1 - 50\hat{\beta}_4 = -13.26341 - 50 \times 0.25684 = -26.1054$

and

$$\text{Var}(\hat{\text{y}}_1 - \hat{\text{y}}_2) = \text{Var}(\hat{\beta}_1) + 50^2\text{Var}(\hat{\beta}_4) + 2 \times 50\text{Cov}(\hat{\beta}_1, \hat{\beta}_4)$$

$$= 60.93273 + 2500 \times 0.027851 + 100 \times (-1.17725) = 12.83523$$

Thus a 95% confidence interval for $\mu_1 - \mu_2$ is: $-26.1054 \pm \sqrt{12.83523} \times t_{.025,84}$

$$= (-26.1054 \pm 1.99 \times 3.58263) = (-33.235, -18.976)$$

This question is based on an observational study of the prevalence of the bacterium *E. coli* in Iowa agricultural fields. One hundred fields were randomly sampled from an enumeration of all fields in the state of Iowa. Researchers visited each selected field and collected 15 soil specimens. Although the locations of these soil specimens were arbitrary, you may assume the 15 soil specimens from any field are a simple random sample of all possible soil specimens within that field. The presence or absence of *E. coli* was assessed in each soil specimen.

Fields were classified by two characteristics: the crop grown on the field (2 levels: corn or soybeans) and the type of manure applied to the field in the spring (3 levels: none, pig, or chicken). Farmers apply manure because it is a source of nitrogen (essential for corn, but less so for soybeans) that is cheaper than inorganic sources. Applying manure to farm fields also provides pig and chicken producers with a way to dispose of unwanted manure. However, manure is a common source of *E. coli*.

The number of fields for each combination of levels of crop and manure is in Table 1.

		Type of Manure		
		None	Pig	Chicken
Crop	Corn	2	44	23
	Soybean	25	3	3

Table 1: Counts of Field Types

Part I

One response variable that can be computed for each field is the number of soil specimens in which *E. coli* is detected. This response is an integer that ranges from 0 to 15 and will be called “number of positive results” in the remainder of Part I. The questions in Part I concern a two-way Analysis of Variance (ANOVA) of the number of positive results. R results from a model fit to these data start on page 6. In all output, the default R parameterization of ‘set first to zero’ for factors and interactions was used. Let β be the vector of parameters associated with the full-rank model matrix determined by this default parameterization. When answering the questions in Part I, please assume the order of the elements of β matches the order of the estimated coefficients on page 6.

1. Find the difference between the estimated marginal mean number of positive results for corn (LSMEAN for corn) and the estimated marginal mean number of positive results for soybean (LSMEAN for soybean).

2. Test the hypothesis that says that the marginal mean number of positive results for

corn is equal to the marginal mean number of positive results for soybean. Report the value of the test statistic, state the distribution of the test statistic under the null hypothesis, and give the p-value.

3. Write the C vector or C matrix so that the null hypothesis of equal marginal means across all three levels of manure can be written as $H_0 : C\beta = 0$.

4. The biologist running the study asks whether the difference in mean response between corn and soybean is the same for all three manure types. Provide a C vector or C matrix so that a test of $H_0 : C\beta = 0$ will address the biologist's question.

5. Using any relevant part of the R code and output on pages 6 and 7, test the hypothesis in question **4**.

6. State the assumptions of the model fit to the data by the R code on page 6.

7. Explain which of the assumptions stated in question **6** are likely to be satisfied and which are likely to be violated when fitting the model to the data by the R code on page 6.

8. Let Y_{ijk} be the number of positive results for crop i, manure type j, and field k. Propose a model for the Y_{ijk} data that is more appropriate than the model fit to these data by the R code on page 6. Define all terms in your proposed model.

Part II

Sampling of specimens was repeated in two years (2012 and 2013) on the same 100 fields. For each field, the same crop and type of manure used in the first year were used again in the second year, and 15 soil specimens were sampled each year. Rather than simply determining the presence or absence of *E. coli* in each specimen, researchers obtained an *E. coli* count for each specimen that is positively associated with *E. coli* abundance. These *E. coli* counts can range from zero to several hundred for different soil specimens. For each combination of field and year, the 15 specimen-specific *E. coli* counts were summed to obtain a response referred to as "total *E. coli* count." This total *E. coli* count will serve as the response variable for the remainder of this problem.

Let Y_{ijkl} be the total *E. coli* count for crop i, manure type j, field k, and year l. Suppose

$$\begin{aligned} Y_{ijkl} &= \mu + \alpha_i + \beta_j + \gamma_l + \nu_{ijk} + \varepsilon_{ijkl} && (1) \\ \nu_{ijk} &\sim N(0, \sigma^2_\nu) && (2) \\ \varepsilon_{ijkl} &\sim N(0, \sigma^2_\varepsilon) && (3) \end{aligned}$$

where all normally distributed random terms are mutually independent and all terms whose distribution is not specified are fixed unknown parameters.

9. Find the correlation between the responses in years 1 and 2 on any single field. Express your answer in terms of the parameters in equations (1) – (3).

10. Suppose the REML method will be used to estimate σ^2_ν and σ^2_ε. The REML method involves finding maximum likelihood estimates of variance parameters using linear combinations of the responses as data. Each of these linear combinations is known as an error contrast. How many error contrasts will be used to compute the REML estimates of σ^2_ν and σ^2_ε in this case?

11. Suppose a statistician questions the analysis of these data using the model in equations (1) – (3). He says that a first-order autoregressive correlation structure [i.e., ar(1)] should be considered for the data from each field because these data are obtained by repeated measures of the response at multiple times. How would you respond to this criticism?

Part III

Now consider the following model.

$$\begin{aligned} \text{Model A :} \qquad Y_{ijkl} \mid \lambda_{ijkl} &\sim \text{Poisson}(\lambda_{ijkl}) && (4) \\ \log \lambda_{ijkl} &= \mu + \alpha_i + \beta_j + \gamma_l + \nu_{ijk} && (5) \\ \nu_{ijk} &\sim N(0, \sigma_\nu^2) && (6) \end{aligned}$$

where the Y_{ijkl} responses are conditionally independent given the λ_{ijkl} terms, all normally distributed random effects are mutually independent, and all terms whose distribution is not specified are fixed unknown parameters.

12. To complete the parts **a**), **b**), and **c**) below, it may help to know the following result concerning a lognormal random variable.

> *If W is a random variable with distribution defined by* $\log W \sim N(\mu, \sigma^2)$, *then*
>
> $$\begin{aligned} \text{E}(W) &= \exp(\mu + \sigma^2/2) \quad \textit{and} \\ \text{Var}(W) &= \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2). \end{aligned}$$

a) Find $\text{E}(Y_{ijkl})$ under Model A. Express your answer in terms of Model A parameters in equations (4) – (6).

b) Find $\text{Var}(Y_{ijkl})$ under Model A. Express your answer in terms of Model A parameters in equations (4) – (6).

c) Find $\text{Cov}(Y_{ijk1}, Y_{ijk2})$ under Model A. Express your answer in terms of Model A parameters in equations (4) – (6).

13. Now consider a new model with both field random effects and observation-specific random effects defined as

$$\text{Model B:} \qquad \begin{aligned} Y_{ijkl} \mid \lambda_{ijkl} &\sim \text{Poisson}(\lambda_{ijkl}) \\ \log \lambda_{ijkl} &= \mu + \alpha_i + \beta_j + \gamma_l + \nu_{ijk} + \tau_{ijkl} \\ \nu_{ijk} &\sim N(0, \sigma^2_\nu) \\ \tau_{ijkl} &\sim N(0, \sigma^2_\tau) \end{aligned}$$

where the Y_{ijkl} responses are conditionally independent given the λ_{ijkl} terms, all normally distributed random effects are mutually independent, and all terms whose distribution is not specified are fixed unknown parameters. Assume Model B holds, and use any relevant part of the R code and output from fitting Model B (found on page 8) to complete parts **a**), **b**), and **c**) below.

a) Compute an estimate that appropriately fills in the blank in the following sentence:

> For any given crop and year, the mean response for fields treated with chicken manure is estimated to be __________ times higher than the mean response for fields treated with no manure.

b) Find an approximate 95% confidence interval for the multiplicative effect estimated in **13 a**).

c) Consider a new field that is not among the 100 previously considered but, like those 100 fields, was randomly sampled from all fields in Iowa. Suppose that corn was planted in this new field and that pig manure was applied. Suppose that, as in the other 100 fields, 15 soil specimens were taken from the new field each year. Let ν be the random field effect for this new field, and let τ_1 be the random observation-specific effect for this new field in year 1 (2012). Let Y_1 be the total *E. coli* count from year 1 (2012) for the new field. According to Model B, the conditional distribution of Y_1, given ν and τ_1, is Poisson. Let λ_1 be the mean of this Poisson distribution.

Suppose that all soil specimens from this new field were lost before any total *E. coli* counts could be determined (i.e., neither Y_1 nor the total *E. coli* count for the new field in year 2 was observed). Using the results from fitting Model B to data from the other 100 fields, find an approximate 95% prediction interval for λ_1.

R code for Part I:

```
Ecoli.lm1 <- lm(Ecoli ~ crop*manure)
summary(Ecoli.lm1)
anova(Ecoli.lm1)
vcov(Ecoli.lm1)
```

R output for Part I:

```
> Ecoli.lm1 <- lm(Ecoli ~ crop*manure)

> summary(Ecoli.lm1)

Call:
lm(formula = Ecoli ~ crop * manure)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9318 -0.9318 -0.1600  1.0682  4.5217

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.4783     0.3550  21.068  < 2e-16 ***
cropSoybean             -0.4783     1.0450  -0.458    0.648
manureNone              -6.9783     1.2549  -5.561 2.51e-07 ***
manurePig                3.4536     0.4380   7.885 5.57e-12 ***
cropSoybean:manureNone   0.1383     1.6300   0.085    0.933
cropSoybean:manurePig   -1.4536     1.4573  -0.997    0.321


Residual standard error: 1.702 on 94 degrees of freedom
Multiple R-squared: 0.8775,     Adjusted R-squared: 0.871
F-statistic: 134.7 on 5 and 94 DF,  p-value: < 2.2e-16
```

```
> anova(Ecoli.lm1)
Analysis of Variance Table

Response: Ecoli
            Df  Sum Sq Mean Sq  F value Pr(>F)
crop         1 1301.65 1301.65 449.1828 <2e-16 ***
manure       2  645.63  322.81 111.3993 <2e-16 ***
crop:manure  2    3.97    1.98   0.6846 0.5068
Residuals   94  272.39    2.90

> vcov(Ecoli.lm1)
                      (Intercept) cropSoybean manureNone  manurePig
(Intercept)             0.1259919  -0.1259919 -0.1259919 -0.1259919
cropSoybean            -0.1259919   1.0919302  0.1259919  0.1259919
manureNone             -0.1259919   0.1259919  1.5748993  0.1259919
manurePig              -0.1259919   0.1259919  0.1259919  0.1918514
cropSoybean:manureNone  0.1259919  -1.0919302 -1.5748993 -0.1259919
cropSoybean:manurePig   0.1259919  -1.0919302 -0.1259919 -0.1918514
                      cropSoybean:manureNone cropSoybean:manurePig
(Intercept)                        0.1259919             0.1259919
cropSoybean                       -1.0919302            -1.0919302
manureNone                        -1.5748993            -0.1259919
manurePig                         -0.1259919            -0.1918514
cropSoybean:manureNone             2.6567501             1.0919302
cropSoybean:manurePig              1.0919302             2.1237279
```

R code for Part III:

```
# full is the data frame containing E. coli counts
# for two years on the same fields

library(lme4)
# model B (both field and observation-specific random effects)
full.glmmB <- glmer(count ~ crop + manure + year +
  (1|field) + (1|obs),
  data=full, family=poisson)

> full.glmmB
Generalized linear mixed model fit by the Laplace approximation
Formula: count ~ crop + manure + year + (1 | field) + (1 | obs)
   Data: full
   AIC   BIC logLik deviance
 475.5 498.6 -230.8    461.5
Random effects:
 Groups Name        Variance Std.Dev.
 obs    (Intercept) 0.042904 0.20713
 field  (Intercept) 0.063961 0.25291
Number of obs: 200, groups: obs, 200; field, 100

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.52875    0.06746   52.31  < 2e-16 ***
cropSoybean -0.06974    0.12145   -0.57    0.566
manureNone  -0.67632    0.13535   -5.00 5.82e-07 ***
manurePig    0.35508    0.07858    4.52 6.22e-06 ***
year2013    -0.48726    0.04067  -11.98  < 2e-16 ***

Correlation of Fixed Effects:
            (Intr) crpSyb manrNn manrPg
cropSoybean -0.209
manureNone  -0.290 -0.727
manurePig   -0.776  0.084  0.317
year2013    -0.267 -0.004  0.007 -0.004
```

Part I

1. The cell means in terms of the default R parameterization are:

	Type of manure		
Crop	None	Pig	Chicken
Corn	$\beta_0 + \beta_n$	$\beta_0 + \beta_p$	β_0
Soybean	$\beta_0 + \beta_n + \beta_s + \beta_{s,n}$	$\beta_0 + \beta_s + \beta_p + \beta_{s,p}$	$\beta_0 + \beta_s$

where β_0 is the intercept, β_n is the coefficient for Manure=none, β_p is the coefficient for Manure=pig, β_s is the coefficient for Crop = soybean, β_s, n is the interaction coefficient for soybean / none, and β_s, p is the interaction coefficient for soybean / pig. Hence, the estimated cell means are:

	Type of manure			Marginal
Crop	None	Pig	Chicken	Mean
Corn	0.50	10.93	7.48	6.30
Soybean	0.16	9.00	7.00	5.39

The marginal (LSMEANS) are the unweighted averages of the cell means.

2. The difference between corn and soybean, each averaged over the three manure types, is estimated by $-\hat{\beta}_s - (\hat{\beta}_{s,n} + \hat{\beta}_{s,p})/3$. The variance of the estimate is

$$\text{Var } \hat{\beta}_s + \frac{\text{Var } \hat{\beta}_{s,n}}{9} + \frac{\text{Var } \hat{\beta}_{s,p}}{9} + \frac{\text{Cov } \hat{\beta}_s,\ \hat{\beta}_{s,n}}{3} + \frac{\text{Cov } \hat{\beta}_s,\ \hat{\beta}_{s,p}}{3} + \frac{\text{Cov } \hat{\beta}_{s,n},\ \hat{\beta}_{s,p}}{9}$$

This variance is estimated by 0.41. Hence, the test statistic is $T = \frac{0.91-0}{\sqrt{0.41}} = 1.43$. Under the null hypothesis, this has a central T distribution with 100 - 6 = 94 degrees of freedom. The exact p-value is 0.156, but tables probably only give you > 0.10.

3. This is a 2 df hypothesis, so the C matrix has 2 rows.

Coefficient					
β_0	β_s	β_n	β_p	$\beta_{s,n}$	$\beta_{s,p}$
0	0	1	-1	0.5	-0.5
0	0	1	0	0.5	0.0

4. The biologist's question can be addressed by the 2 df test of the interaction between crop and manure. Any constant times the coefficients in the table below would yield an appropriate C matrix.

Coefficient					
β_0	β_s	β_n	β_p	$\beta_{s,n}$	$\beta_{s,p}$
0	0	0	0	0	-1
0	0	0	0	-1	0

5. You could construct a $C\beta$ test, but the anova table contains a valid test of interaction: F = 0.68, central F with 2,94 df, p = 0.51.

6. The responses are independent. Each response has a normal distribution. For each response, the mean of its normal distribution may depend on the combination of crop and manure type applied to the field, but the variance is the same for all responses.

7. The independence assumption is reasonable if we assume simple random sampling of 100 fields from a large population of fields. The response (number of positive results) cannot be normally distributed because it takes integer values in $\{0, 1, 2, \ldots, 15\}$. A binomial distribution, as described in the answer to the next question, would be more appropriate. If responses follow binomial distributions, the mean of each response may depend on the combination of crop and manure type applied to the field, but the variance would not necessarily be the same for all responses. For example, we would expect fields with very low or very high *E. coli* abundance to have less variable responses than fields with moderate levels of *E. coli* abundance due to the relationship between mean and variance for binomial distributions.

8. The responses are likely to have binomial distributions, for which an appropriate model is:

$$\begin{aligned} Y_{ijk} &\sim \text{Binomial}(15, \pi_{ij}) \\ \log \frac{\pi_{ij}}{1-\pi_{ij}} &= \mu + \alpha_i + \beta_j + \alpha\beta_{ij}, \end{aligned}$$

where α_i is a crop-specific contribution to the log odds, β_j is a manure-type-specific contribution, and $\alpha\beta_{ij}$ is an interaction (cell-specific) contribution. Other models are possible, including one with additional variability for fields.

Part II

9. The correlation between two observations from the same field is $\frac{\sigma^2_\nu}{\sigma^2_\nu+\sigma^2_\varepsilon}$.

10. The number of error contrasts is the total sample size ($n = 200$) minus the rank of the model matrix that defines the mean of the response vector ($r = 4$). Thus, the number of error contrasts is 196.

11. The correlation structure for two observations from any field in the model with field random effects implies a correlation structure that is equivalent to an ar(1) correlation structure. Because there are two observations per field, there is only one parameter for the correlation. The ar(1) coefficient, i.e., the correlation between two observations from the same field is $\frac{\sigma_\nu^2}{\sigma_\nu^2+\sigma_\varepsilon^2}$.

Part III

12. To simplify notation, let $\mu_{ijl} = \mu + \alpha_i + \beta_j + \gamma_l$.

a)

$$\begin{aligned} \mathrm{E}(Y_{ijkl}) &= \mathrm{E}\left\{\mathrm{E}(Y_{ijkl}|\lambda_{ijkl})\right\} = \mathrm{E}\left\{\exp(\mu_{ijl}+\nu_{ijk})\right\} \\ &= \exp(\mu_{ijl}+\sigma_\nu^2/2) \end{aligned}$$

b)

$$\begin{aligned} \mathrm{Var}(Y_{ijkl}) &= \mathrm{Var}\left\{\mathrm{E}(Y_{ijkl}|\lambda_{ijkl})\right\} + \mathrm{E}\left\{\mathrm{Var}(Y_{ijkl}|\lambda_{ijkl})\right\} \\ &= \mathrm{Var}\left\{\exp(\mu_{ijl}+\nu_{ijk})\right\} + \mathrm{E}\left\{\exp(\mu_{ijl}+\nu_{ijk})\right\} \\ &= \exp(2\mu_{ijl}+2\sigma_\nu^2) - \exp(2\mu_{ijl}+\sigma_\nu^2) + \exp(\mu_{ijl}+\sigma_\nu^2/2) \end{aligned}$$

c)

$$\begin{aligned} \mathrm{Cov}(Y_{ijk1},Y_{ijk2}) &= \mathrm{E}(Y_{ijk1}Y_{ijk2}) - \mathrm{E}(Y_{ijk1})\mathrm{E}(Y_{ijk2}) \\ &= \mathrm{E}\left\{\mathrm{E}(Y_{ijk1}Y_{ijk2}|\nu_{ijk})\right\} - \exp(\mu_{ij1}+\sigma_\nu^2/2)\exp(\mu_{ij2}+\sigma_\nu^2/2) \\ &= \mathrm{E}\left\{\exp(\mu_{ij1}+\nu_{ijk})\exp(\mu_{ij2}+\nu_{ijk})\right\} - \exp(\mu_{ij1}+\mu_{ij2}+\sigma_\nu^2) \\ &= \mathrm{E}\left\{\exp(\mu_{ij1}+\mu_{ij2}+2\nu_{ijk})\right\} - \exp(\mu_{ij1}+\mu_{ij2}+\sigma_\nu^2) \\ &= \exp(\mu_{ij1}+\mu_{ij2}+2\sigma_\nu^2) - \exp(\mu_{ij1}+\mu_{ij2}+\sigma_\nu^2) \end{aligned}$$

13. **a**) $\exp(0.67632) \approx 1.97$

b) $\exp(0.67632 - 2 * 0.13535)$ to $\exp(0.67632 + 2 * 0.13535)$, i.e., 1.50 to 2.58.

c) We seek an approximate 95% prediction interval for

$$\lambda_1 = \exp(\mu + \alpha_{\mathrm{corn}} + \beta_{\mathrm{pig}} + \gamma_1 + \nu + \tau_1).$$

The linear combination of parameters

$$\mu + \alpha_{\mathrm{corn}} + \beta_{\mathrm{pig}} + \gamma_1 \text{ is estimated as } 3.52875 + 0 + 0.35508 + 0 = 3.88383.$$

The estimated variance associated with this estimate is

$$0.06746^2 + 0.07858^2 - 2 * 0.776 * 0.06746 * 0.07858 = 0.002498505.$$

An approximate 95% prediction interval for $\mu + \alpha_{\text{corn}} + \beta_{\text{pig}} + \gamma_1 + \nu + \tau_1$ is

$$3.88383 \pm 2\sqrt{0.002498505 + 0.063961 + 0.042904} \Longleftrightarrow (3.222427, 4.545233).$$

The desired prediction interval for λ_1 is then approximately

$$(\exp(3.222427), \exp(4.545233)) \Longleftrightarrow (25.1, 94.2).$$

The number of problems that involve temporally structured data have increased as data collection technologies have become more sophisticated (such as those based on automated sensors) and as computer simulation models become more widely used in areas such as climate and weather forecasting. This question concerns statistical structures that might be used in modeling such data. To gain a feel for the types of problems under consideration, we will briefly consider two examples.

The Iowa Department of Transportation (DOT) has an automated system for collecting data on traffic speed and related variables at a number of locations on highways in the state. The primary purpose of these sensors is to record traffic speeds during winter storm events, which are then used in assessing the efficacy of winter storm surface maintainence. A plot of traffic speeds recorded by one of these sensors, and averaged to 5 minute periods for a particular snow storm at a particular location, is presented in Figure 1 (Figures begin on page 6). The DOT currently has a number of sensors that record data for roughly 25 to 40 event/location combinations in a winter.

Another example of the type of problems of concern is long-range forecasting of maximum daily temperatures in July. These forecasts are made beginning in February. For each month from February to June, a numerical weather model is initialized at four different times on each of three days, resulting in a total of 60 model runs (for a number of locations in Iowa). The only factors that vary in these model runs are the initial conditions at which the model is started. The weather model is, in fact, deterministic so that if the model is started with the same values twice it will return exactly the same temperature forecasts. Despite this, the model is so complex that the internal dynamics of the interacting differential equations that make up the model are not understood, and it has become common practice to model the outputs of these models as if they arose from an underlying stochastic process. Figure 2 presents a plot of forecasted maximum daily temperatures in July for a location near Ames, Iowa,for two model runs started on 25 February 1982, one started at hour 00 and the other started at hour 18.

These two examples share several characteristics. One is that an assumption of constant (marginal) mean is motivated by the underlying problem. Weather models are provided no information about average or “typical” temperatures on the various days of July. With suitable covariates we might attempt to model a portion of the variability in the traffic speed data in terms of those covariates, but none are available. A second characteristic of these two problems is that there is a definitive starting point ($t = 0$) for each. That is, observation cannot be thought of as beginning at an arbitrary point for a process that has been in operation for a long period of time. These two characteristics will be important in our consideration of models for temporal structure. That there is non-trivial temporal structure is shown in the autocorrelation and partial autocorrelation functions of Figure 3 for the traffic speed data and Figure 4 for the long-range temperature forecasts.

ANSWER QUESTION 1 NOW

(Questions begin on page 13.)

Part I: Dynamic Models for Individual Time Sequences

Consider a temporal sequence of random variables $\boldsymbol{Y} = \{Y(t) : t = 0, 1, \dots n\}$. We assume that these variables correspond to an observable quantity of interest (e.g., traffic speeds or temperatures) made at equally spaced points in time. Our objective is to formulate models that might be used to represent $\boldsymbol{Y}$. Two characteristics that we wish any such model to possess are that they should (1) result in constant marginal means and (2) result in temporal correlation that decreases as random variables become more greatly separated in time. One approach for formulating such models is to make use of dynamic model structures.

Consider a potential model, for $t = 1, 2, \dots, n$ specified by

$$\begin{aligned} Y(t) &= \mu(t) + \epsilon_t \\ \mu(t) &= \lambda + \gamma\,[\mu(t-1) - \lambda] \end{aligned} \qquad (1)$$

where the ϵ_t are independent and identically distributed having normal distributions with mean 0 and variance σ^2, $-1 < \gamma < 1$, $-\infty < \lambda < \infty$, $\sigma^2 > 0$, independent of

$$\mu(0) \sim N(\lambda, \tau^2).$$

ANSWER QUESTION 2 NOW

(Questions begin on page 13.)

An alternative to model (1) is, for $t = 1, \ldots, n$ specified by

$$\begin{aligned} Y(t) &= \mu(t) + \epsilon_t \\ \mu(t) &= \lambda + \gamma\left[\mu(t-1) - \lambda\right] + v_t, \end{aligned} \tag{2}$$

where the ϵ_t are independent and identically distributed having normal distributions with mean 0 and variance σ^2 and are independent of the v_t that are independent and identically distributed having normal distributions with mean 0 and variance τ^2, $-1 < \gamma < 1$, $-\infty < \lambda < \infty$, $\sigma^2 > 0$, $\tau^2 > 0$, all independent of

$$\mu(0) \sim N\left(\lambda, \frac{\tau^2}{(1-\gamma^2)}\right).$$

ANSWER QUESTION 3 NOW

(Questions begin on page 13.)

Yet another possible model having dynamic structure is, for $t = 1, \ldots, n$ specified by

$$\begin{aligned} Y(t) &= \mu(t) + \epsilon_t \\ \mu(t) &= \mu(t-1) + v_t, \end{aligned} \tag{3}$$

where the ϵ_t are independent and identically distributed having normal distributions with mean 0 and variance σ^2, and are independent of the v_t that are independent and identically distributed having normal distributions with mean 0 and variance τ^2, $-\infty < \lambda < \infty$, $\sigma^2 > 0$, $\tau^2 > 0$, all independent of

$$\mu(0) \sim N\left(\lambda, \tau^2\right).$$

ANSWER QUESTION 4 NOW

(Questions begin on page 13.)

Because of the problem with model (1) you identified in Question 2 we will not further consider model (1). Two simulated data sets of length 100 are presented for model (2) in the upper row of Figure 5 and two simulated data sets of length 100 are presented for model (3) in the lower row of Figure 5. In all of these simulated examples, $\lambda = 75$, $\sigma^2 = 2.0$, $\tau^2 = 0.5$ and (for model (2) only) $\gamma = 0.70$.

The realizations from models (2) and (3) shown in Figure 5 should exhibit the behaviors you identified in Question 3 and Question 4 and these behaviors are evident given that we can view multiple realizations of each model. In many applications, however, we would have only one series of data to analyze, and it might not be clear from an exploratory plot which model we would prefer to fit. For example, it is not obvious which model generated the realization shown in Figure 6, which was produced with the same parameter values as used for the realizations of Figure 5.

ANSWER QUESTIONS 5 and 6 NOW

(Questions begin on page 13.)

Part II: Autoregressive Errors for Individual Time Sequences

An alternative to the dynamic models of Part I is to model data using additive errors with autoregressive structure,

$$\begin{aligned} Y(t) &= \mu + W(t) \\ W(t) &= \gamma W(t-1) + \epsilon_t, \end{aligned} \tag{4}$$

where the ϵ_t are independent and identically distributed random variables having normal distributions with mean 0 and variance $\sigma^2 > 0$ independent of

$$W(0) \sim N\left(0,\, \frac{\sigma^2}{(1-\gamma^2)}\right).$$

For this model

$$\begin{aligned} E[Y(t)] &= \lambda \\ var[Y(t)] &= \frac{\sigma^2}{1-\gamma^2} \\ cov[Y(t),\, Y(t+k)] &= \frac{\gamma^k \sigma^2}{1-\gamma^2} \\ cor[Y(t),\, Y(t+k)] &= \gamma^k \end{aligned}$$

So model (4) possesses the two characteristics we desire and, like model (2), is second order stationary.

ANSWER QUESTIONS 7 and 8 NOW

(Questions begin on page 13.)

A composite likelihood was constructed for use in estimation of the parameters of model (4). A set of data was simulated from model (4) using parameter values $\mu = 75$, $\sigma^2 = 2$ and $\gamma = 0.7$. Making use of the log composite likelihood with the R functions `optim` and `nlm` resulted in "convergence" of estimates to different values from different starting values, although the convergence codes returned by these functions indicated no problems. To investigate this problem, slices of the log composite likelihood were calculated. While one could compute any number of such slices, four particular ones are presented in Figure 7. The upper left plot of Figure 7 represents a slice in the dimension of γ for $\mu = 50$ and $\sigma^2 = 2$. The upper right plot of Figure 7 represents a slice in the dimension of σ^2 for $\mu = 50$ and $\gamma = 0.7$. The lower left plot of Figure 7 represents a slice in the dimension of μ for $\sigma^2 = 8$ and $\gamma = 0.7$, and the lower right plot represents a slice in the dimension of γ for $\mu = 75$ and $\sigma^2 = 8$.

ANSWER QUESTIONS 9, 10 and 11 NOW

(Questions begin on page 13)

Figures

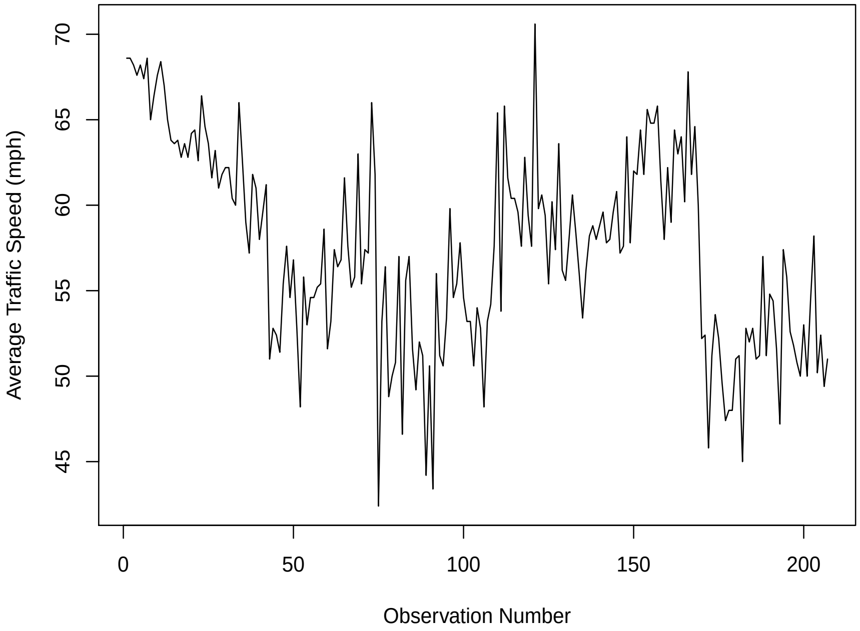


Figure 1: Plot of recorded traffic speeds at a particular location in Iowa during a winter storm event. Observation number indexes average values across 5 minute intervals with 1 indicating the beginning of the storm event.

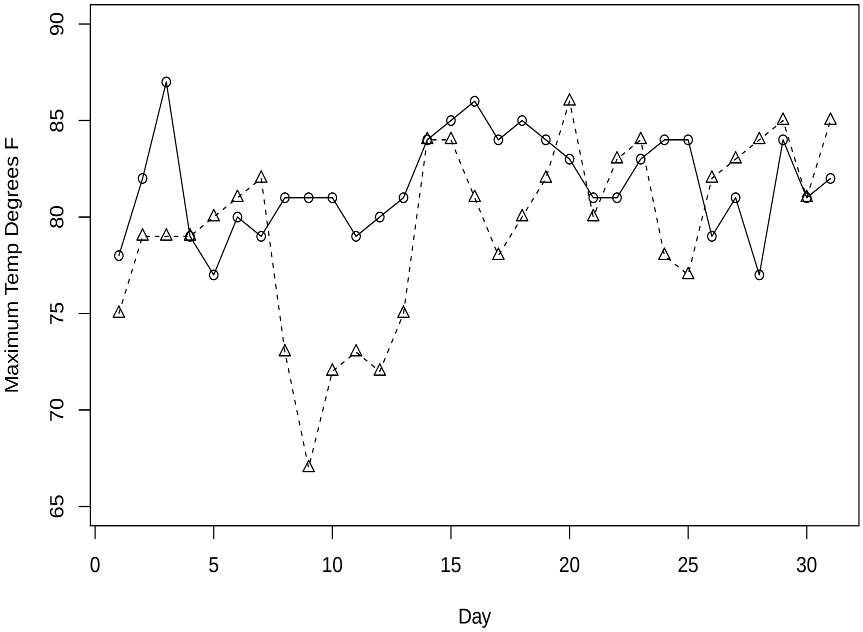


Figure 2: Long-range forecasts of maximum daily temperatures in July for a location near Ames, Iowa. The solid line depicts forecasts from a model started at hour 00 and the dashed line a model started at hour 18, both on 25 February 1982. (Hours here are in terms of Coordinated Universal Time or what is called Zulu Time by the military.)

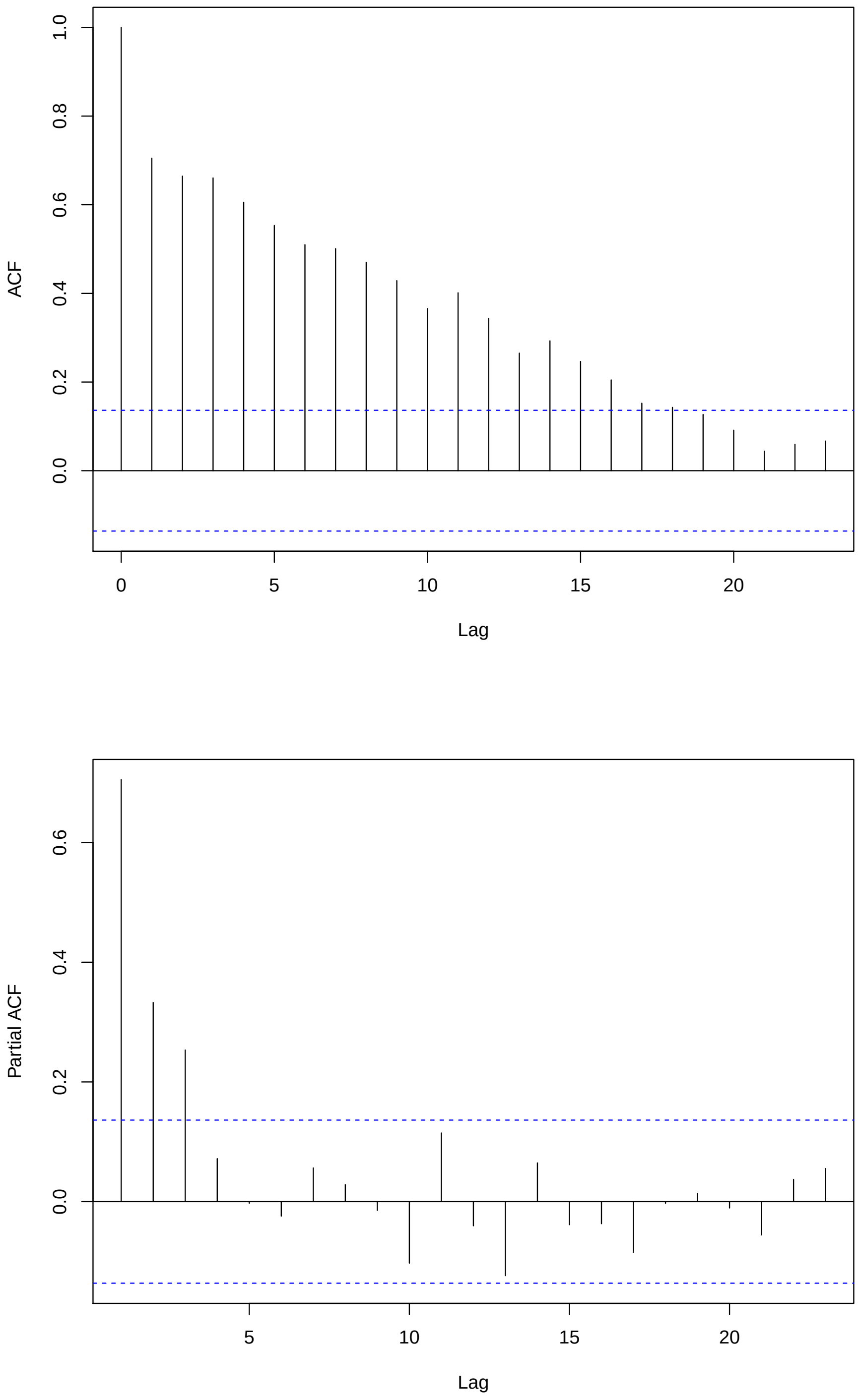


Figure 3: Autocorrelation and partial autocorrelation for traffic speed data.

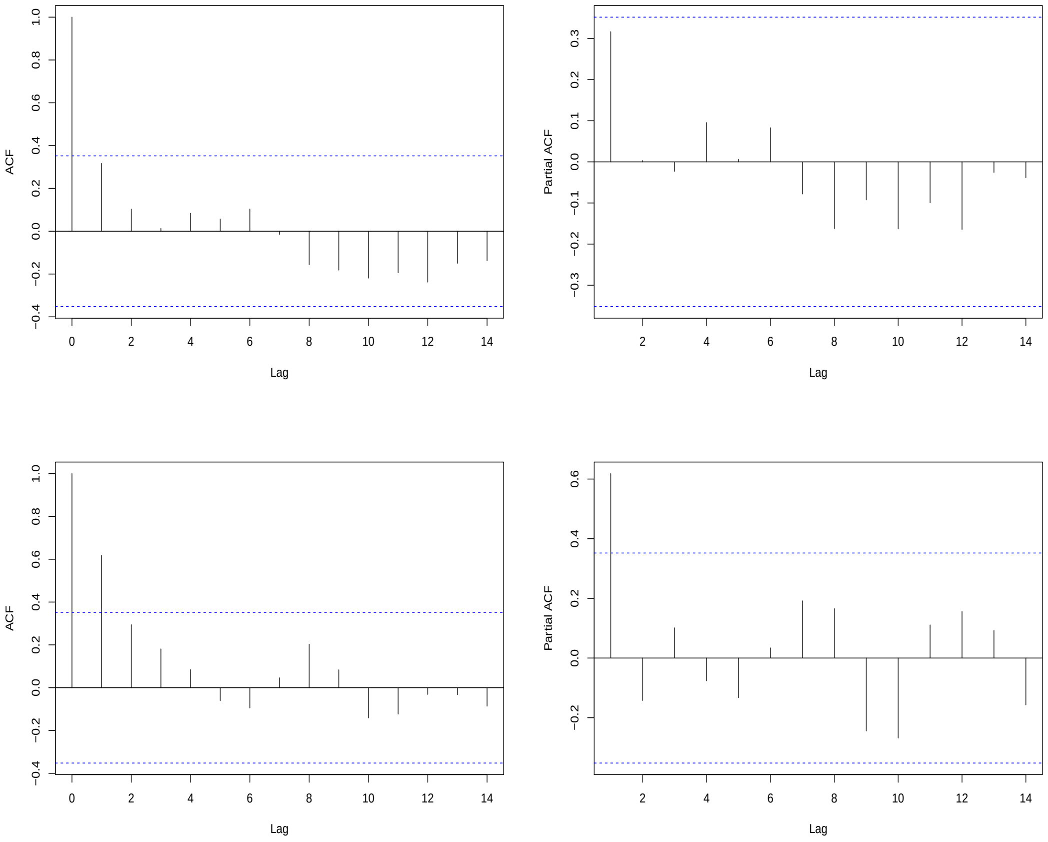


Figure 4: Autocorrelation and partial autocorrelation for long-range temperature forecasts. The top two plots are for time 00 and the lower two plots are for time 18.

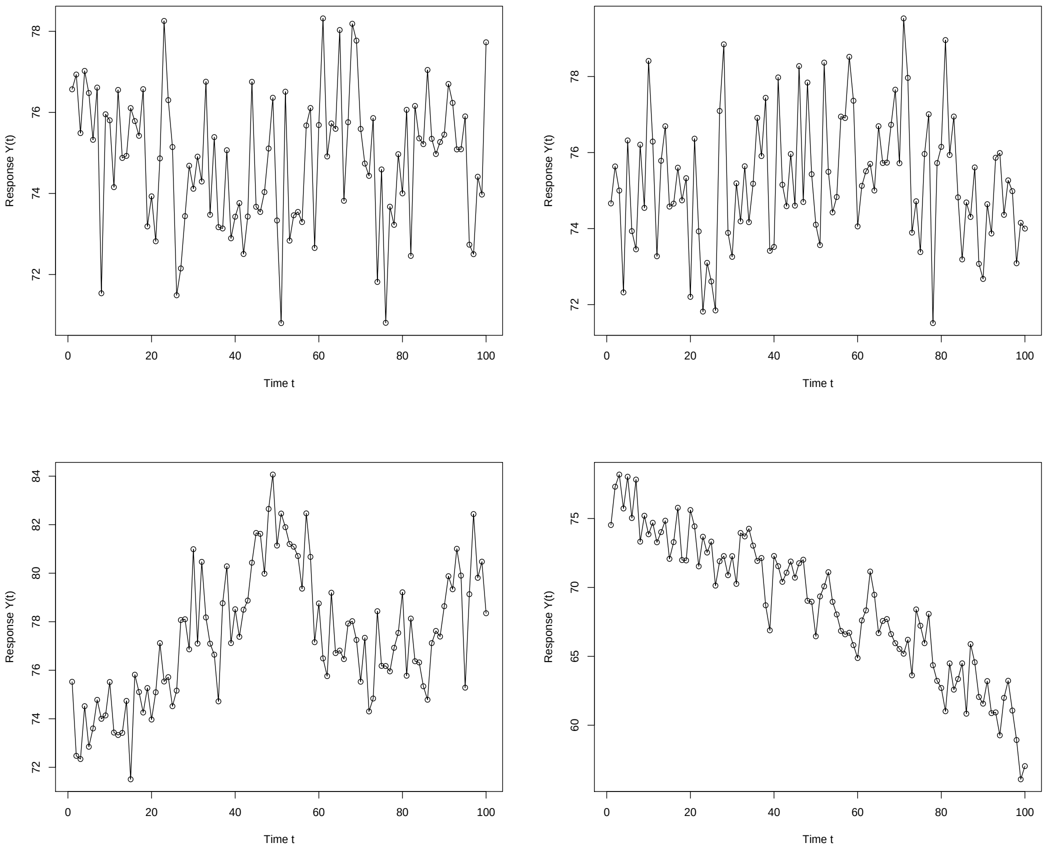


Figure 5: Two realizations from model (2) in the upper row and two realizations from model (3) in the lower row. Parameter values were $\lambda = 75$, $\sigma^2 = 2.0$, $\tau^2 = 0.5$, and $\gamma = 0.70$.

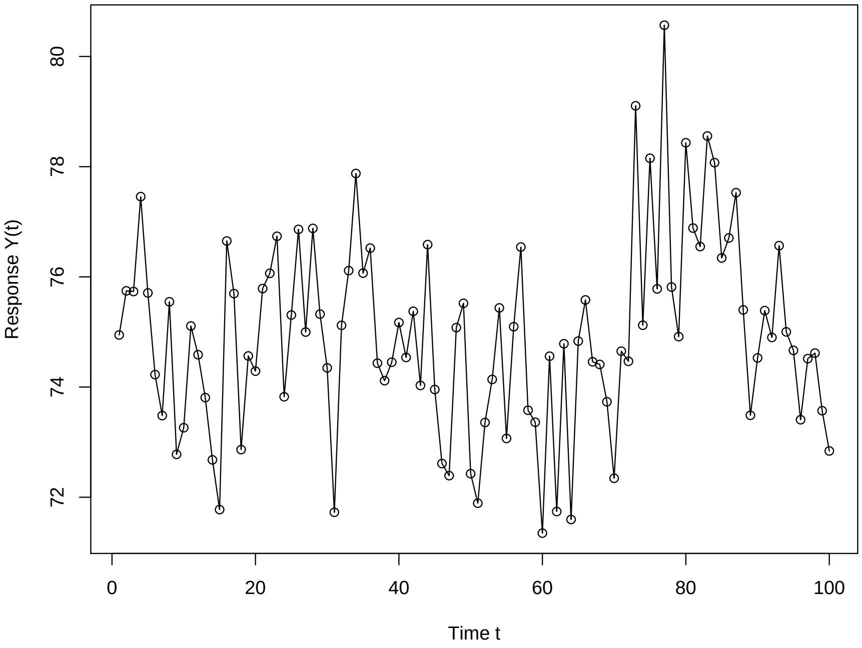


Figure 6: A simulated data set for which the identity of the generating model is not obvious.

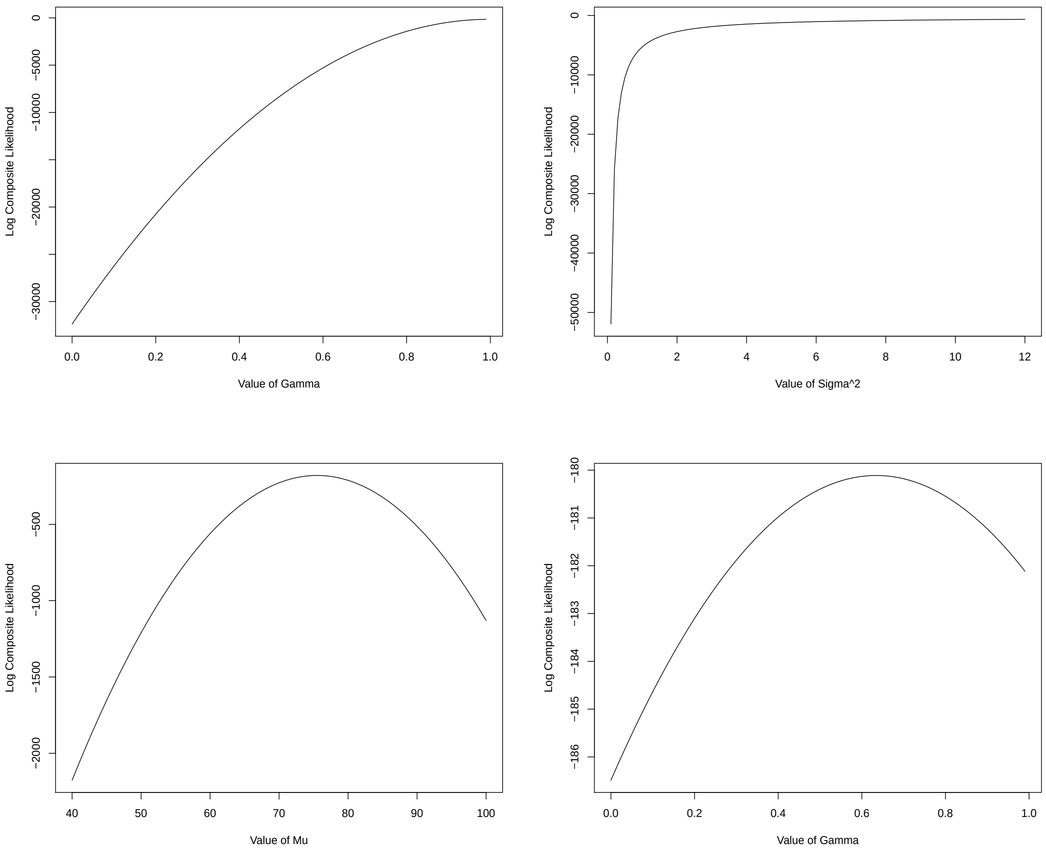


Figure 7: Slices of a composite log likelihood for fitting model (4) to a set of simulated data.

Questions

1. In all of the models we will consider we will assume a constant marginal mean, so that the models will differ in the way that variances and statistical dependencies are represented in small-scale model structures. Explain what additional issue in model formulation would complicate our ability to compare models if we were not able to make this assumption of a constant mean.

2. Demonstrate that (with a suitable restriction on allowable values of γ) model (1) possesses the two characteristics of constant mean and correlation that decreases with time that we desire in our models. Do this by deriving expected values, variances, covariances, and correlations. Demonstrate what happens to variances as time increases without bound. Explain why this implies that model (1) would not be useful for data such as those represented in Figures 1 and 2.

3. Demonstrate that (with a suitable restriction on allowable values of γ) model (2) possesses the two characteristics (constant mean and correlation that decreases with time) that we desire in our models. Demonstrate additionally that this model is second order stationary and thus has positive variances for all points in time.

4. Demonstrate that model (3) possesses the two characteristics of constant mean and correlation that decreases with time that we desire in our models. Demonstrate additionally that, although this model is not second order stationary, like model (2) it does have positive variances for all points in time.

5. Explain (one or two sentences are all that is needed) why models specified by (2) are not in the class of models specified by (3).

6. Suppose that both model (2) and model (3) are fit to a given set of data, such as those represented in Figure 6. Outline, in algorithmic form, a simulation-based procedure that could be used to choose between these two models.

   Hint: Keep in mind that both models have a constant marginal mean.

7. Models (2) and (4) are quite similar. Suppose we are interested in developing either a simulation-based procedure that might distinguish between these models (similar to the procedure you provided in answer to Question 6) or a data-driven diagnostic that can be used with a single set of data without relying on simulation. In either case, success will depend on whether you can identify an aspect or characteristic of

data behavior that should differ between these two models and determine a way to quantify that characteristic. Based on quantities you computed to answer previous questions suggest a characteristic of data behavior that might have promise in our development, then suggest a way to quantify the characteristic you have selected. *Hint: There is a sample autocovariance function as well as a sample autocorrelation function. The autocovariance function provides estimates of the covariances of random variables separated by time lags $k = 0, 1, 2, \ldots$. Two data sets were simulated, one from Model 2 and one from Model 4 with parameters chosen to match expected values and variances of the models as closely as possible. The first five autocovariances for these simulated data are as follows:*

Model	k=0	k=1	k=2	k=3	k=4
2	4.1685	1.0804	0.9291	0.7138	0.3939
4	2.2008	1.6005	1.2019	0.8013	0.4340

8. Although it would be possible to derive the full likelihood for model (4), this is a situation in which using a composite likelihood might be attractive. Give a careful development of a composite likelihood appropriate for this model.

   Hint: The phrase "careful development" in this question implies that you should define a set of either marginal or conditional "events" (and the corresponding probabilities) on which to base your composite likelihood.

9. Although not exhaustive of the various manners by which one might examine the log composite likelihood surface, what procedure for determining parameter estimates is suggested by the log composite likelihood slices shown in Figure 7?

10. One approach for making inferences (such as computing interval estimates) would be to rely on asymptotic normality for maximum composite likelihood estimators, if we can find a suitable theoretical result that covers our particular application. For the problem of forecasting maximum daily temperatures in July, describe the asymptotic context within which we should search for one or more appropriate results.

11. As an alternative to estimation of a limiting covariance matrix to be used in an application of asymptotic normality (e.g., Wald theory), we could form interval estimates directly through application of a parametric bootstrap. Explain why the existence of an asymptotic result is important for this bootstrap procedure, even if it is not used explicitly in computations.

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 issue in question is that there is no unique manner to partition data structure into what are often called *large scale* and *small scale* model components. Large scale model structure is connected with marginal expectations while small scale model structure is connected with variances and dependencies. Thus, if there was non-constant marginal mean structure, the manner in which that structure was modeled could influence how the small-scale model would be best dealt with, thus complicating the comparison of models in terms of how they represent statistical dependencies.

Question 2. Beginning with $E[\mu(0)] = \lambda$ we have that

$$\begin{aligned} E[\mu(1)] &= E[\lambda + \gamma\,\{\mu(0) - \lambda\}] = \lambda \\ E[\mu(2)] &= E[\lambda + \gamma\,\{\mu(1) - \lambda\}] = \lambda \end{aligned}$$

It follows by induction that, for $t = 1, \ldots, n$,

$$E[\mu(t)] = \lambda.$$

This model thus implies a constant marginal mean, as desired. In the same way, beginning with $var[\mu(0)] = \tau^2$,

$$\begin{aligned} var[\mu(1)] &= var[\lambda + \gamma\,\{\mu(0) - \lambda\}] = \gamma^2\,\tau^2 \\ var[\mu(2)] &= var[\lambda + \gamma\,\{\mu(1) - \lambda\}] = \gamma^4\,\tau^2 \end{aligned}$$

It follows by induction that, for $t = 1, \ldots, n$,

$$var[\mu(t)] = \gamma^{2t}\,\tau^2.$$

Given $E[\mu(t)] = \lambda$ for all t,

$$\begin{aligned} cov[\,u(t), \mu(t+1)] &= E[\mu(t)\mu(t+1)] - \lambda^2 \\ &= E[\mu(t)\{\lambda + \gamma\,\mu(t) - \gamma\,\lambda\}] - \lambda^2 \\ &= \lambda E[\mu(t)] + \gamma\,E[\mu^2(t)] - \gamma\lambda E[\mu(t)] - \lambda^2 \\ &= \gamma\,\Big(var[\mu(t)] + \lambda^2\Big) - \gamma\lambda^2 \\ &= \gamma\gamma^{2t}\tau^2 = \gamma^{2t+1}\tau^2. \end{aligned}$$

Similarly,

$$\begin{aligned} cov[\,u(t), \mu(t+2)] &= E[\mu(t)\mu(t+2)] - \lambda^2 \\ &= E[\mu(t)\{\lambda + \gamma\,\mu(t+1) - \gamma\,\lambda\} - \lambda^2 \\ &= \lambda E[\mu(t)] + \gamma\,E[\mu(t)\mu(t+1)] - \gamma\lambda E[\mu(t)] - \lambda^2 \\ &= \gamma\,\Big(cov[\mu(t), \mu(t+1)] + \lambda^2\Big) - \gamma\lambda^2 \\ &= \gamma\gamma^{2t+1}\tau^2 = \gamma^{2t+2}\tau^2. \end{aligned}$$

Continuing in this fashion shows that

$$cov[\mu(t), \mu(t+k)] = \gamma^{2t+k}\tau^2.$$

With $Y(t) = \mu(t) + \epsilon_t$ with ϵ_t independent and identically distributed $N(0, \sigma^2)$,

$$\begin{aligned} E[Y(t)] &= E[\mu(t)] = \lambda \\ var[Y(t)] &= var[\mu(t)] + var[\epsilon_t] = \gamma^{2t}\tau^2 + \sigma^2 \\ cov[Y(t), Y(t+k)] &= cov[\mu(t), \mu(t+k)] = \gamma^{2t+k}\tau^2 \end{aligned}$$

As a result, $Y(t)$ has constant expected value and

$$\begin{aligned} cor[Y(t), (t+k)] &= \frac{\gamma^{2t+k}\tau^2}{\left(\{\gamma^{2t}\tau^2 + \sigma^2\}\{\gamma^{2t+k}\tau^2 + \sigma^2\}\right)^{1/2}} \\ &< \frac{\gamma^{2t+k}\tau^2}{\left(\{\gamma^{2t}\tau^2\}\{\gamma^{2t+k}\tau^2\}\right)^{1/2}} \\ &< \frac{\gamma^{2t+k}\tau^2}{\left(\{\gamma^{2t+k}\gamma^{-k}\tau^2\}\{\gamma^{2t+k}\tau^2\}\right)^{1/2}} \\ &< \gamma^k. \end{aligned}$$

If $|\gamma| < 1$ then correlations will decay with increasing separation in time, as is desired.

While the model implies constant marginal mean and correlations that decay with time, the variance $var[\mu(t)] = \gamma^{2t}\tau^2$ will go to zero if $|\gamma| < 1$. This severely limits the usefulness of the model in that, for any substantial time record, the model implies the response variables $Y(t)$ are essentially independent and identically distributed according to a normal distribution with mean λ and variance σ^2. That is, as the variance of the sequence $\mu(t)$ approaches zero, the model approaches $Y(t) = \lambda + \epsilon_t$.

Question 3. Beginning with $E[\mu(0)] = \lambda$ and $E[v_t] = 0$ for all t we have that

$$\begin{aligned} E[\mu(1)] &= E[\lambda + \gamma\,\{\mu(0) - \lambda\} + v_t] = \lambda \\ E[\mu(2)] &= E[\lambda + \gamma\,\{\mu(1) - \lambda\} + v_t] = \lambda \end{aligned}$$

It follows by induction that, for $t = 1, \ldots, n$,

$$E[\mu(t)] = \lambda.$$

For the variances,

$$\begin{aligned} var[\mu(1)] &= var[\lambda + \gamma\,\{\mu(0) - \lambda\} + v_t] \\ &= \gamma^2 var[\mu(0)] + \tau^2 = \frac{\gamma^2\tau^2}{1-\gamma^2} + \tau^2 = \frac{\tau^2}{1-\gamma^2} \\ var[\mu(2)] &= var[\lambda + \gamma\,\{\mu(1) - \lambda\} + v_t] \\ &= \gamma^2 var[\mu(1)] + \tau^2 = \frac{\gamma^2\tau^2}{1-\gamma^2} + \tau^2 = \frac{\tau^2}{1-\gamma^2} \end{aligned}$$

It follows by induction that, for $t = 1, \ldots, n$,

$$var[\mu(t)] = \frac{\tau^2}{1-\gamma^2}.$$

With $E[\mu(t)] = \lambda$ for $t = 1, \ldots, n$,

$$\begin{aligned} cov[\mu(t),\, \mu(t+1)] &= E[\mu(t)\mu(t+1)] - \lambda^2 \\ &= E[\mu(t)\{\lambda + \gamma\,\mu(t) - \gamma\,\lambda + v_t\}] - \lambda^2 \end{aligned}$$

$$\begin{aligned}
&= \lambda E[\mu(t)] + \gamma\, E[\mu^2(t)] - \gamma\lambda E[\mu(t)] + E[\mu(t)v_t] - \lambda^2 \\
&= \gamma\, \left(var[\mu(t)] + \lambda^2\right) - \gamma\lambda^2 \\
&= \frac{\gamma\tau^2}{1-\gamma^2}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
cov[\,u(t), \mu(t+2)] &= E[\mu(t)\mu(t+2)] - \lambda^2 \\
&= E[\mu(t)\{\lambda + \gamma\,\mu(t+1) - \gamma\,\lambda\} - \lambda^2 \\
&= \lambda E[\mu(t)] + \gamma\, E[\mu(t)\mu(t+1)] - \gamma\lambda E[\mu(t)] + E[\mu(t)v_t] - \lambda^2 \\
&= \gamma\, \left(cov[\mu(t), \mu(t+1)] + \lambda^2\right) - \gamma\lambda^2 \\
&= \frac{\gamma^2\tau^2}{1-\gamma^2}.
\end{aligned}$$

Then by induction,

$$cov[\mu(t), \mu(t+k)] = \frac{\gamma^k\tau^2}{1-\gamma^2}.$$

Then with $Y(t) = \mu(t) + \epsilon_t$, $\epsilon_t \sim iidN(0, \sigma^2)$, and $k \neq 0$

$$\begin{aligned}
E[Y(t)] &= E[\mu(t)] = \lambda \\
var[Y(t)] &= var[\mu(t)] + var[\epsilon_t] = \frac{\tau^2}{1-\gamma^2} + \sigma^2 \\
cov[Y(t), Y(t+k)] &= cov[\mu(t), \mu(t+k)] = \frac{\gamma^k\tau^2}{1-\gamma^2}
\end{aligned}$$

Then assuming that $|\gamma| < 1$, $Y(t)$ has constant expected value, constant variance, covariance that depends only on time separation and

$$\begin{aligned}
cor[Y(t), (t+k)] &= \left[\frac{\gamma^k\tau^2}{1-\gamma^2}\right] / \left[\frac{\tau^2}{1-\gamma^2} + \sigma^2\right] \\
&< \left[\frac{\gamma^k\tau^2}{1-\gamma^2}\right] / \left[\frac{\tau^2}{1-\gamma^2}\right] = \gamma^k, \qquad (1)
\end{aligned}$$

which depends only on time lag k, and decays with increasing k for $|\gamma| < 1$. Thus, the stochastic process $\{Y(t) : t = 0, 1, \ldots\}$ has the two characteristics we are seeking and is second order stationary.

Question 4. For this model $E[\mu(0)] = \lambda$ and

$$\begin{aligned} E[\mu(1)] &= E[\{\mu(0) + v_t] = \lambda \\ E[\mu(2)] &= E[\mu(1) + v_t] = \lambda \end{aligned}$$

It follows by induction that, for $t = 1, \ldots, n$,

$$E[\mu(t)] = \lambda.$$

With $var[\mu(0)] = \tau^2$,

$$\begin{aligned} var[\mu(1)] &= var[\mu(0) + v_t] = 2\,\tau^2 \\ var[\mu(2)] &= var[\mu(1) + v_t] = 3\,\tau^2 \end{aligned}$$

It follows then that

$$var[\mu(t)] = (t+1)\tau^2.$$

With $E[\mu(t)] = \lambda$ for $t = 1, \ldots, n$,

$$\begin{aligned} cov[\mu(t),\, \mu(t+1)] &= E[\mu(t)\mu(t+1)] - \lambda^2 \\ &= E[\mu(t)\{\mu(t) + v_t\}] - \lambda^2 \\ &= E[\mu^2(t)] + E[\mu(t)v_t] - \lambda^2 \\ &= var[\mu(t)] + \lambda^2 - \lambda^2 = var[\mu(t)] = (t+1)\tau^2 \end{aligned}$$

Similarly,

$$\begin{aligned} cov[\,u(t), \mu(t+2)] &= E[\mu(t)\mu(t+2)] - \lambda^2 \\ &= E[\mu(t)\{\mu(t+1) + v_t\}] - \lambda^2 \\ &= E[\mu(t)\mu(t+1)] + E[\mu(t)v_t] - \lambda^2 \\ &= cov[\mu(t),\, \mu(t+1)] = var[\mu(t)] = (t+1)\tau^2 \end{aligned}$$

Then by induction,

$$cov[\mu(t),\, \mu(t+k)] = var[\mu(t)] = (t+1)\tau^2.$$

Using the model $Y(t) = \mu(t) + \epsilon_t$, $\epsilon_t \sim iidN(0, \sigma^2)$, and for $k \neq 0$,

$$\begin{aligned} E[Y(t)] &= E[\mu(t)] = \lambda \\ var[Y(t)] &= var[\mu(t)] + var[\epsilon_t] = (t+1)\tau^2 + \sigma^2 \\ cov[Y(t), Y(t+k)] &= cov[\mu(t), \mu(t+k)] = (t+1)\tau^2 \end{aligned}$$

so that

$$cor[Y(t), Y(t+k)] = \frac{(t+1)\tau^2}{\left[\{(t+1)\tau^2 + \sigma^2\}\{(t+k+1)\tau^2 + \sigma^2\}\right]^{1/2}},$$

which, for a fixed t, is a decreasing function of k.

As a result, this model has constant marginal mean and correlation that decreases with time, as desired. While the variances must be positive (for $t \geq 0$) they are not constant, so this model is not second order stationary. Interestingly, for a given t, covariance is constant in time, and it is the increase in variance of $Y(t+k)$ over $Y(t)$ that causes the correlation to decrease as k grows.

Question 5. While we could obtain model (3) from model (2) by setting $\gamma = 1$, this value is not in the parameter space for model (2) which is $-1 < \gamma < 1$.

Question 6. Both of these models imply a constant marginal mean but, as illustrated in Figure 5, the stationary model (2) exhibits more regular behavior in fluctuating around that mean, while the random walk incorporated into model (3) causes realizations to meander or wander so that longer portions of the sequence lie entirely above or entirely below the marginal mean. These behaviors can be quantified by counting how often a realized squence of length n crosses the average of values in the sequence. This can be quantified as

$$\bar{Y} = \frac{1}{n}\sum_{t=1}^{n} Y(t),$$

$$Q(\boldsymbol{Y}) = \sum_{t=1}^{n} \left(I[Y(t+1) > \bar{Y} | Y(t) \leq \bar{Y}] + I[Y(t+1) \leq \bar{Y} | Y(t) > \bar{Y}] \right),$$

where $I(x)$ is the indicator function that assumes a value of 1 if x is true and a value of 0 otherwise.

For a set of observed data $\boldsymbol{y} = \{y(t) : t = 1, \ldots, n\}$ this quantity can be computed by defining sets

$$\begin{aligned} B &= \{y(t) : y(t) \leq \bar{y}\} \\ A &= \{y(t) : y(t) < \bar{y}\} \end{aligned}$$

and then

$$Q(\boldsymbol{y}) = \sum_{y(t) \in B} I[y(t+1) > \bar{y}] + \sum_{y(t) \in A} I[(t+1) \leq \bar{y}], \qquad (2)$$

where $I(x)$ is the indicator function as previously.

Given estimated parameter values we can easily simulate values from either model (2) or model (3) and a procedure to assess either of these models is given by the following algorithm.

(a) Simulate data sets $\boldsymbol{y}_m^*$, each of size n from a fitted model, $m = 1, \ldots, M$.

(b) For $m = 1, \ldots, M$, compute $\bar{y}_m^*$ and $Q_m^* = Q(\boldsymbol{y}_m^*)$ where this is given by expression (1) in this solution.

(c) For $\boldsymbol{y}$ denoting the actual data, compute $\bar{y}$ and $Q^a = Q(\boldsymbol{y})$,

(d) A simulation-based $p-$value for the ability of the model to reflect the identified aspect of data behavior is then

$$\begin{aligned} p_R &= \sum_{m=1}^{M} I[Q_m^* \geq Q^a] \\ p_L &= \sum_{m=1}^{M} I[Q_m^* \leq Q^a] \\ p &= \min\{p_R, p_L\}. \end{aligned}$$

A small value of p would indicate that the model under assessment is not fully adequate at reflecting the identified behavior.

Question 7. Previous derivations and results given in the question give the quantities listed in Table 1 (for $k \neq 0$). Models (2) and (4) could be "matched" to a large degree though parameter relations such as taking τ^2 in Model (2) to have the same value as σ^2 in Model (4).

Quantity	Model (2)	Model (4)
$E[Y(t)]$	λ	λ
$var[Y(t)]$	$\frac{\tau^2}{1-\gamma^2}+\sigma^2$	$\frac{\sigma^2}{1-\gamma^2}$
$cov[Y(t), Y(t+k)]$	$\frac{\gamma^k\tau^2}{1-\gamma^2}$	$\frac{\gamma^k\sigma^2}{1-\gamma^2}$
$cor[Y(t), Y(t+k)]$	$\frac{\gamma^k\tau^2}{\tau^2+\sigma^2(1-\gamma^2)}$	γ^k

Table 1: Moments of Model (2) and Model (4).

Assuming $\gamma > 0$, covariances for both models at time lags of $k = 1, 2, \ldots$ should decay as a power of some number between 0 and 1. A difference between these models is that for a given time lag k the covariance of Model (4) is directly proportional to variance while for Model (2) covariance is proportional to variance with an additional offset. That is, in Model (4), $cov[Y(t), Y(t+k)] = \gamma^k var[Y(t)]$ while in Model (2), $cov[Y(t), Y(t+k)] = \gamma^k var[Y(t)] - \gamma^k\sigma^2$. In particular, under Model (2) $var[Y(t)] = \gamma cov[Y(t), Y(t+1)] + \sigma^2$ while under Model (4) $var[Y(t)] = \gamma cov[Y(t), Y(t+1)]$. Therefore, a potential diagnostic might be based on whether autocovariance at lag $k = 0$ (i.e., variance) fits into the pattern of decay for those at greater lags or whether it contains an additional “spike”.

To quantify this characteristic of data behavior, consider the forms of the covariances given in Table 1. Choose some value of M as the greatest lag at which covariances are still meaningful. For either model, and for $t = 2, \ldots, M-1$

$$\frac{cov[Y(t), Y(t+k)]}{cov[Y(t), Y(t+k-1)]} = \gamma$$

Now, for Model (4),

$$\frac{cov[Y(t), Y(t+1)]}{cov[Y(t), Y(t)]} = \frac{cov[Y(t), Y(t+1)]}{var[Y(t)]} = \gamma.$$

But for Model (2)

$$\frac{cov[Y(t),Y(t+1)]}{cov[Y(t),Y(t)]} = \frac{cov[Y(t),Y(t+1)]}{var[Y(t)]} = \frac{\gamma\tau^2}{\tau^2+\sigma^2(1-\gamma^2)} < \gamma.$$

Then compute

$$\begin{aligned} Q_1 &= \frac{1}{M-1}\sum_{k=2}^{M} \frac{\hat{cov}[Y(t),Y(t+k)]}{\hat{cov}[Y(t),Y(t+k-1)]} \\ Q_0 &= \frac{\hat{cov}[Y(t),Y(t+1)]}{\hat{cov}[Y(t),Y(t)]} = \frac{\hat{cov}[Y(t),Y(t+1)]}{\hat{var}[Y(t)]} \\ D &= Q_1 - Q_0 \\ D^* &= Q_1/Q_0. \end{aligned}$$

The quantity D (or, alternatively, D^*) embodies the characteristic of data behavior desired. Larger values of D (or D^* substantially greater than 1) suggest data behavior more in line with Model (2) and smaller values of D (or D^* near 1) suggest data behavior more in line with Model (4).

For the values provided in the Hint to this question,

Quantity	Model (2)	Model (4)
Q_1	0.7267	0.6531
Q_0	0.2592	0.7272
D	0.4675	-0.0741
D^*	2.804	0.898

Question 8. The model directly provides the conditional densities, for $t = 2, \ldots, n$,

$$f(y(t)|y(t-1),\mu,\sigma^2,\gamma) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left\{y(t)-\gamma y(t-1)-\mu(1-\gamma)\right\}^2\right].$$

Let k index pairs $(t, t-1)$; $t = 2, \ldots, n$ so that $k = 1$ corresponds to $(2, 1)$ and $k = n-1$ corresponds to $(n, n-1)$. Let the events A_k, $k = 1, \ldots, n-1$ be defined for some small δ as

$$y(t)-\delta < Y(t) < y(t)+\delta \text{ given that } y(t-1)-\delta < Y(t-1) < y(t-1)+\delta$$

A composite likelihood can then be defined for these events as

$$\tilde{L_c}(\mu,\sigma^2,\gamma) = \prod_{i=1}^{n-1} L_k(\mu,\sigma^2,\gamma) = \prod_{i=1}^{n-1} Pr(A_k),$$

or, using density approximations,

$$L_c(\mu,\sigma^2,\gamma) = \prod_{t=2}^{n} f(y(t)|y(t-1),\mu,\sigma^2,\gamma),$$

with $f(\cdot)$ given at the start of this answer.

Question 9. These slices of the log composite likelihood surface suggest profiling out σ^2. It appears that the composite likelihood might be nicely behaved in the dimensions of μ and γ when σ^2 is not near its estimate. This suggests the profile log composite likelihood

$$\ell_c^p(\sigma^2) = \max_{\mu,\gamma} \ell_c(\mu,\sigma^2,\gamma).$$

Maximizing $\ell_c^p(\sigma^2)$ in σ^2 then provides the simultaneous maximum composite likelihood estimators.

Question 10. The most appropriate asymptotic context for the problem of forecasting maximum daily temperatures in July would be that of "repeating lattices". This can be seen to be the case because there are always exactly 31 random variables in any realized stochastic process. With the number of random variables fixed, our context must be that of obtaining multiple realizations, rather than allowing a single realization to grow large (expanding lattice).

Question 11. In a parametric bootstrap procedure we are approximating the sampling distribution of some function of a parameter and an estimator of that parameter for a fixed sample size. We need some assurance that this distribution exists. If the function of parameter and estimator being bootstrapped has a limit distribution we have obtained this assurance and we can be confident we are approximating at some point in a sequence of distributions. Since any finite collection of numbers defines an empirical distribution we cannot rely on computer output alone to provide the justification needed.