PhD Prelim Exam
METHODS

Summer 2011
(Given on 7/5/11)

Exposure to lead can have adverse health effects, especially in young children. Lead poisoning in the United States occurs mainly through exposure to old paint containing lead. Children who have been exposed to lead can be treated with drugs that help excretion of the lead.

A new drug called *succimer* is said to promote excretion more effectively than existing treatments. A placebo-controlled randomized trial called the *TLC trial* (for *Treatment of Lead-exposed Children*) was conducted to determine whether treatment with succimer reduces blood lead levels over time more than what would be observed when treating children with placebo.

One hundred children with confirmed blood lead levels of 20 - 44 μg/dL (micrograms of lead per one deciliter of blood) aged 12 - 33 months at enrollment were allocated to a placebo or a succimer group using a completely randomized design with 50 subjects in each group. Children received a treatment (either succimer or placebo) at baseline (week 0), and their blood lead levels were then measured at weeks 1 and 6 post-treatment. Children were followed for three years but we consider only the initial three measurements here. We use N to denote the number of children in the study and n to denote the number of measurements made on each child. Therefore, $N = 100$ and $n = 3$.

Table 1 shows an excerpt of the complete dataset. Table 2 shows the means by week for the placebo and succimer groups. Figure 1 is a plot of the means over time. Each mean in Table 2 and Figure 1 is an average of 50 datapoints.

ID	Group	Baseline	Week 1	Week 6
79	P (Placebo)	31	27	24
8	S (Succimer)	27	15	21
44	S	26	23	23
11	P	25	25	23
69	S	20	3	9
29	S	20	5	12
46	P	29	21	18
13	P	34	32	25
74	P	20	15	15
53	P	31	31	30
⋮	⋮	⋮	⋮	⋮

Table 1: *Blood lead levels (μg/dL) per time point for children from the TLC trial.*

Group	Baseline	Week 1	Week 6
Succimer	27	14	21
Placebo	26	25	24

Table 2: *Mean blood lead levels ($\mu g/dL$) per time point for children from the TLC trial.*

We use Y_{ijk} to denote the blood lead measurement on the kth child in the ith group at the jth time point, with $i = P, S$, $j = 1,\ 2,\ 3$ and $k = 1, ..., 50$. Measurement occasions are denoted by t so that $t_1 = 0$, $t_2 = 1$ and $t_3 = 6$ weeks. For each child, we can define a 3×1 vector $\mathbf{Y}_{ik}$ of measurements. We let

$$\mathbf{Y}_{ik} = \begin{bmatrix} Y_{i1k} \\ Y_{i2k} \\ Y_{i3k} \end{bmatrix}, \quad \boldsymbol{\mu}_i = E(\mathbf{Y}_{ik}) = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i3} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = Var(\mathbf{Y}_{ik}) = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \sigma_{i13} \\ & \sigma_{i22} & \sigma_{i23} \\ & & \sigma_{i33} \end{bmatrix}. \tag{1}$$

The sample covariance matrices of the response vectors in the succimer (S) and placebo (P) groups are denoted as $\mathbf{V}_S$ and $\mathbf{V}_P$ respectively, and are given below.

$$\mathbf{V}_S = \begin{bmatrix} 25.2 & 15.4 & 23.2 \\ 15.4 & 58.8 & 36.2 \\ 23.2 & 36.2 & 85.6 \end{bmatrix}, \text{ and } \mathbf{V}_P = \begin{bmatrix} 25.2 & 22.8 & 21.5 \\ 22.8 & 29.8 & 23.4 \\ 21.5 & 23.4 & 31.8 \end{bmatrix}. \tag{2}$$

Part I

Given the information you have received so far, please answer the following questions:

1. Are these summary statistics consistent with the claim that children were randomized to groups? Conduct a test or provide a confidence interval to support your answer.

2. Scientists are interested in testing whether changes in mean blood lead levels over time differ across the two treatment groups. They formulated two different null hypotheses. Explain the difference between these two null hypotheses in the context of this study.

$$\text{(A) } H_0 \ : \ \mu_{Sj} = \mu_{Pj}, \quad j = 1, 2, 3$$
$$\text{(B) } H_0 \ : \ \mu_{Sj} - \mu_{S1} = \mu_{Pj} - \mu_{P1}, \quad j = 2, 3.$$

3. Consider only the succimer group and test whether there was a change in mean blood lead levels between baseline ($j = 1$) and week 1 ($j = 2$).

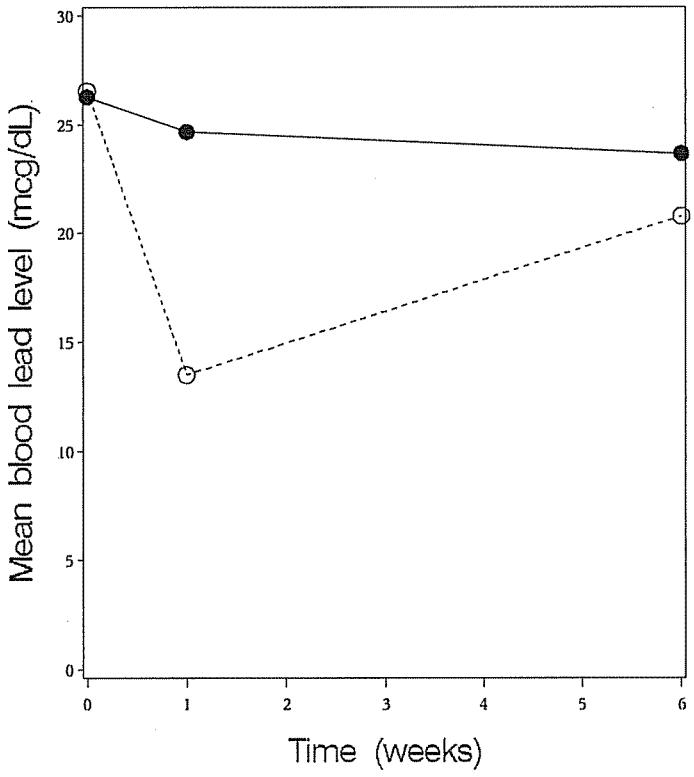


Figure 1: *Mean blood levels ($\mu g/dL$) per time point for children from the TLC trial. Black dots correspond to the placebo group and open circles correspond to the succimer group.*

4. An alternative design for a study with a similar objective consists in randomly allocating 150 children to each treatment group but then measuring a random sub-sample of 50 children in each treatment group at each measuring occasion. In this design, we also obtain a total of 300 measurements, but each child provides only one measurement of blood lead level. Which design (this, or the design used in the *TLC* study) would you prefer and why?

Part II

Refer to Figure 1 and Table 2. An approach to comparing the mean responses over time in the succimer and the placebo groups consists in computing the *areas under the curves* (AUC) between baseline ($t_1 = 0$) and week 6 ($t_3 = 6$). The shaded areas in Figure 2 show the sample AUC for the placebo and the succimer groups on the left and right panels, respectively.

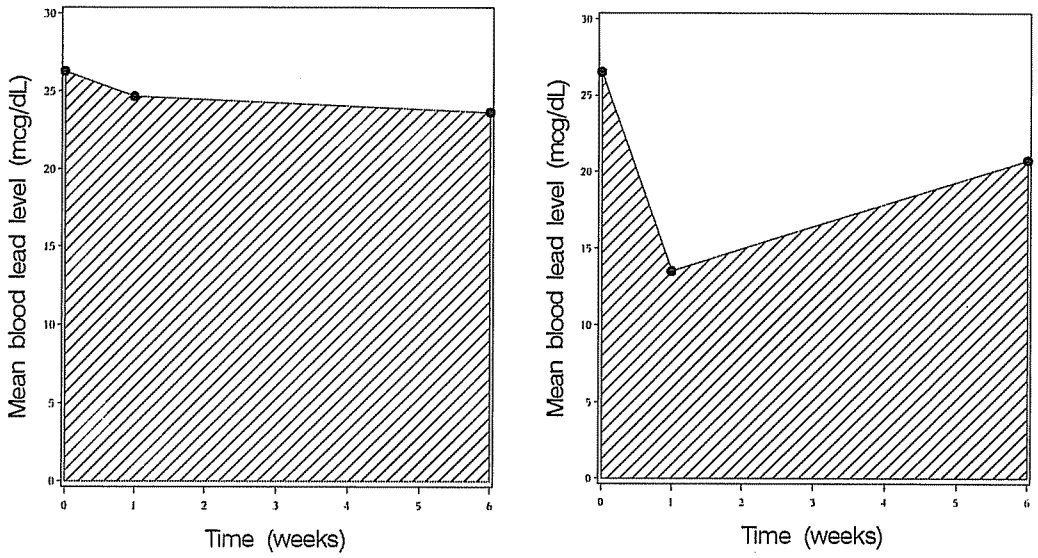


Figure 2: *Area under the response curves. The left panel corresponds to the placebo group and the right panel corresponds to the succimer group.*

5. Let AUC_P and AUC_S denote population versions of the sample areas displayed in the left and right panels of Figure 2, respectively. Conduct a test of the hypothesis that AUC_P equals AUC_S. Assume that the response vectors for all children in a treatment group are 3-dimensional multivariate normal.

Part III

Two models were fit to these data. In both cases, the vector of blood lead levels for the kth subject in the ith treatment group, $\mathbf{Y}_{ik}$, was assumed to be $\mathrm{N}_3(\mathbf{X}_{ik}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\mathbf{X}_{ik}$ a $3 \times p$ design matrix, p the dimension of the vector of unknown regression coefficients $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ the 3×3 covariance matrix. The response vectors for the $N = 100$ children in the study are assumed to be independent. The covariance matrix $\boldsymbol{\Sigma}$ was assumed to be homogeneous across treatment groups. Treatment group is indicated by the dummy variable g_i that takes on values 0 or 1 for placebo and succimer groups respectively. In both models, the covariance matrix $\boldsymbol{\Sigma}$ has a compound symmetric structure, so that

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_3 + \sigma_b^2 \boldsymbol{J}_3,$$

where $\boldsymbol{I}_3$ is a 3×3 identity matrix, $\boldsymbol{J}_3$ is a 3×3 matrix of ones, and σ_b^2, σ^2 are unknown scalar-valued parameters with $\sigma_b^2 > 0$ and $\sigma^2 > 0$.

Model 1 was a linear time trend model with

$$E(Y_{ijk}) = \beta_0 + \beta_1 g_i + \beta_2 t_j + \beta_3 g_i t_j.$$

Model 2 was a linear spline model, with a knot at week 1, so that

$$\begin{aligned} E(Y_{ijk}) &= \beta_0 + \beta_1 g_i + \beta_2 t_j + \beta_3 g_i t_j \\ &\quad + \beta_4 (t_j - 1)_+ + \beta_5 g_i (t_j - 1)_+, \end{aligned}$$

where $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$.

6. Under Model 1 provide an expression for $E(Y_{ijk})$ for all combinations of $i \in \{P, S\}$ and $j = 1, 2, 3$.

7. Under Model 2 provide an expression for $E(Y_{ijk})$ for all combinations of $i \in \{P, S\}$ and $j = 1, 2, 3$.

8. Use the attached SAS code and output to estimate the variance of the following differences under Model 1:

   a) $Y_{ijk} - Y_{ijk'}$, $k \neq k'$

   b) $Y_{ijk} - Y_{ij'k'}$, $j \neq j'$; $k \neq k'$

   c) $Y_{ijk} - Y_{ij'k}$, $j \neq j'$.

9. Refer to the attached SAS code and output. Under Model 1, test the hypothesis that the changes in mean lead blood levels over time in the succimer group are the same as those in the placebo group.

10. Refer to the attached SAS code and output. Under Model 2, test the hypothesis that the changes in mean lead blood levels over time in the succimer group are the same as those in the placebo group.

```
data tlc;
     input id grp $ lead0 lead1 lead6;
     y=lead0; time=0; output;
     y=lead1; time=1; output;
     y=lead6; time=6; output;
     drop lead0 lead1 lead6;
run ;

data tlc ; set tlc ;
  t = time ;
  ctime = time ;
  time_1 = max(time-1,0) ;
  if grp = 'A' then group = 1 ;
  if grp = 'P' then group = 0 ;
  timesqr = time * time ;
run ;

title 'Linear trend model - Compound symmetry cov';
proc mixed data = tlc ;
  class id t ;
  model y = group time group*time / s chisq covb;
  repeated t / type = cs subject = id rcorr ;
run ;

title 'Spline model with knot at week 1 - Compound symmetry cov';
proc mixed data = tlc ;
  class id t ;
  model y = group time group*time time_1 group*time_1 / s chisq covb;
  repeated t / type = CS subject = id rcorr ;
run ;
```

Linear trend model - Compound symmetry cov

The Mixed Procedure

Model Information

Data Set	WORK.TLC
Dependent Variable	y
Covariance Structure	Compound Symmetry
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information

Class	Levels	Values
id	100	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
t	3	0 1 6

Dimensions

Covariance Parameters	2
Columns in X	4
Columns in Z	0
Subjects	100

Max Obs Per Subject 3

Linear trend model - Compound symmetry cov

The Mixed Procedure

Number of Observations

Number of Observations Read	300
Number of Observations Used	300
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	2059.82970025	
1	1	2037.31969779	0.00000000

Convergence criteria met.

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3
1	1.0000	0.2907	0.2907
2	0.2907	1.0000	0.2907
3	0.2907	0.2907	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	id	16.5554
Residual		40.3973

Linear trend model - Compound symmetry cov

The Mixed Procedure

Fit Statistics

-2 Res Log Likelihood	2037.3
AIC (smaller is better)	2041.3
AICC (smaller is better)	2041.4
BIC (smaller is better)	2046.5

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	22.51	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	25.7038	0.9018	98	28.50	<.0001
group	-4.9968	1.2754	98	-3.92	0.0002
time	-0.3619	0.1977	198	-1.83	0.0687
group*time	0.1766	0.2796	198	0.63	0.5283

Covariance Matrix for Fixed Effects

Row	Effect	Col1	Col2	Col3	Col4
1	Intercept	0.8133	-0.8133	-0.09122	0.09122
2	group	-0.8133	1.6265	0.09122	-0.1824
3	time	-0.09122	0.09122	0.03909	-0.03909
4	group*time	0.09122	-0.1824	-0.03909	0.07819

Linear trend model - Compound symmetry cov
The Mixed Procedure

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
group	1	98	15.35	15.35	<.0001	0.0002
time	1	198	3.35	3.35	0.0672	0.0687
group*time	1	198	0.40	0.40	0.5276	0.5283

Spline model with knot at week 1 - Compound symmetry cov
The Mixed Procedure

Model Information

Data Set	WORK.TLC
Dependent Variable	y
Covariance Structure	Compound Symmetry
Subject Effect	id
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Between-Within

Class Level Information

Class	Levels	Values
id	100	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100
t	3	0 1 6

Dimensions

Covariance Parameters	2
Columns in X	6
Columns in Z	0
Subjects	100
Max Obs Per Subject	3

Spline model with knot at week 1 - Compound symmetry cov
The Mixed Procedure

Number of Observations

Number of Observations Read	300
Number of Observations Used	300
Number of Observations Not Used	0

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	1968.26132718	
1	1	1883.22445230	0.00000000

Convergence criteria met.

Estimated R Correlation Matrix for id 1

Row	Col1	Col2	Col3
1	1.0000	0.5536	0.5536
2	0.5536	1.0000	0.5536
3	0.5536	0.5536	1.0000

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
CS	id	23.6614
Residual		19.0795

Spline model with knot at week 1 - Compound symmetry cov
The Mixed Procedure

Fit Statistics

-2 Res Log Likelihood	1883.2
AIC (smaller is better)	1887.2
AICC (smaller is better)	1887.3
BIC (smaller is better)	1892.4

Null Model Likelihood Ratio Test

DF	Chi-Square	Pr > ChiSq
1	85.04	<.0001

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	26.2720	0.9246	98	28.42	<.0001
group	0.2680	1.3075	98	0.20	0.8380
time	-1.6120	0.8736	196	-1.85	0.0665
group*time	-11.4060	1.2355	196	-9.23	<.0001
time_1	1.4092	0.9728	196	1.45	0.1490
group*time_1	13.0568	1.3757	196	9.49	<.0001

Covariance Matrix for Fixed Effects

Row	Effect	Col1	Col2	Col3	Col4	Col5	Col6
1	Intercept	0.8548	-0.8548	-0.3816	0.3816	0.3816	-0.3816
2	group	-0.8548	1.7096	0.3816	-0.7632	-0.3816	0.7632
3	time	-0.3816	0.3816	0.7632	-0.7632	-0.8395	0.8395
4	group*time	0.3816	-0.7632	-0.7632	1.5264	0.8395	-1.6790

Spline model with knot at week 1 - Compound symmetry cov
The Mixed Procedure

Covariance Matrix for Fixed Effects

Row	Effect	Col1	Col2	Col3	Col4	Col5	Col6
5	time_1	0.3816	-0.3816	-0.8395	0.8395	0.9463	-0.9463
6	group*time_1	-0.3816	0.7632	0.8395	-1.6790	-0.9463	1.8927

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	Chi-Square	F Value	Pr > ChiSq	Pr > F
group	1	98	0.04	0.04	0.8376	0.8380
time	1	196	3.40	3.40	0.0650	0.0665
group*time	1	196	85.23	85.23	<.0001	<.0001
time_1	1	196	2.10	2.10	0.1474	0.1490
group*time_1	1	196	90.07	90.07	<.0001	<.0001

Exposure to lead can have adverse health effects, especially in young children. Lead poisoning in the United States occurs mainly through exposure to old paint containing lead. Children who have been exposed to lead can be treated with drugs that help excretion of the lead.

A new drug called *succimer* is said to promote excretion more effectively than existing treatments. A placebo-controlled randomized trial called the *TLC trial* (for *Treatment of Lead-exposed Children*) was conducted to determine whether treatment with succimer reduces blood lead levels over time more than what would be observed when treating children with placebo.

One hundred children with confirmed blood lead levels of 20 - 44 μg/dL (micrograms of lead per one deciliter of blood) aged 12 - 33 months at enrollment were allocated to a placebo or a succimer group using a completely randomized design with 50 subjects in each group. Children received a treatment (either succimer or placebo) at baseline (week 0), and their blood lead levels were then measured at weeks 1 and 6 post-treatment. Children were followed for three years but we consider only the initial three measurements here. We use N to denote the number of children in the study and n to denote the number of measurements made on each child. Therefore, $N = 100$ and $n = 3$.

Table 1 shows an excerpt of the complete dataset. Table 2 shows the means by week for the placebo and succimer groups. Figure 1 is a plot of the means over time. Each mean in Table 2 and Figure 1 is an average of 50 datapoints.

ID	Group	Baseline	Week 1	Week 6
79	P (Placebo)	31	27	24
8	S (Succimer)	27	15	21
44	S	26	23	23
11	P	25	25	23
69	S	20	3	9
29	S	20	5	12
46	P	29	21	18
13	P	34	32	25
74	P	20	15	15
53	P	31	31	30
⋮	⋮	⋮	⋮	⋮

Table 1: *Blood lead levels (μg/dL) per time point for children from the TLC trial.*

Group	Baseline	Week 1	Week 6
Succimer	27	14	21
Placebo	26	25	24

Table 2: *Mean blood lead levels (μg/dL) per time point for children from the TLC trial.*

We use Y_{ijk} to denote the blood lead measurement on the kth child in the ith group at the jth time point, with $i = P, S$, $j = 1,\ 2,\ 3$ and $k = 1, ..., 50$. Measurement occasions are denoted by t so that $t_1 = 0$, $t_2 = 1$ and $t_3 = 6$ weeks. For each child, we can define a 3×1 vector $\mathbf{Y}_{ik}$ of measurements. We let

$$\mathbf{Y}_{ik} = \begin{bmatrix} Y_{i1k} \\ Y_{i2k} \\ Y_{i3k} \end{bmatrix}, \quad \boldsymbol{\mu}_i = E(\mathbf{Y}_{ik}) = \begin{bmatrix} \mu_{i1} \\ \mu_{i2} \\ \mu_{i3} \end{bmatrix}, \quad \boldsymbol{\Sigma}_i = Var(\mathbf{Y}_{ik}) = \begin{bmatrix} \sigma_{i11} & \sigma_{i12} & \sigma_{i13} \\ & \sigma_{i22} & \sigma_{i23} \\ & & \sigma_{i33} \end{bmatrix}. \tag{1}$$

The sample covariance matrices of the response vectors in the succimer (S) and placebo (P) groups are denoted as $\mathbf{V}_S$ and $\mathbf{V}_P$ respectively, and are given below.

$$\mathbf{V}_S = \begin{bmatrix} 25.2 & 15.4 & 23.2 \\ 15.4 & 58.8 & 36.2 \\ 23.2 & 36.2 & 85.6 \end{bmatrix}, \text{ and } \mathbf{V}_P = \begin{bmatrix} 25.2 & 22.8 & 21.5 \\ 22.8 & 29.8 & 23.4 \\ 21.5 & 23.4 & 31.8 \end{bmatrix}. \tag{2}$$

Part I

Given the information you have received so far, please answer the following questions:

1. Are these summary statistics consistent with the claim that children were randomized to groups? Conduct a test or provide a confidence interval to support your answer.

   Yes. The randomization appears to have been effective because at baseline (first measurement occasion, or week 0) the means and the variances of blood lead level are similar for both groups. A simple t-test of the hypothesis that $\mu_{S1} = \mu_{P1}$ results in a statistic with value $t = 1.00/1.00$, clearly not significant.

2. Scientists are interested in testing whether changes in mean blood lead levels over time differ across the two treatment groups. They formulated two different null hypotheses. Explain the difference between these two null hypotheses in the context of this study.

   (A) $H_0 : \mu_{Sj} = \mu_{Pj}, \quad j = 1, 2, 3$

   (B) $H_0 : \mu_{Sj} - \mu_{S1} = \mu_{Pj} - \mu_{P1}, \quad j = 2, 3.$

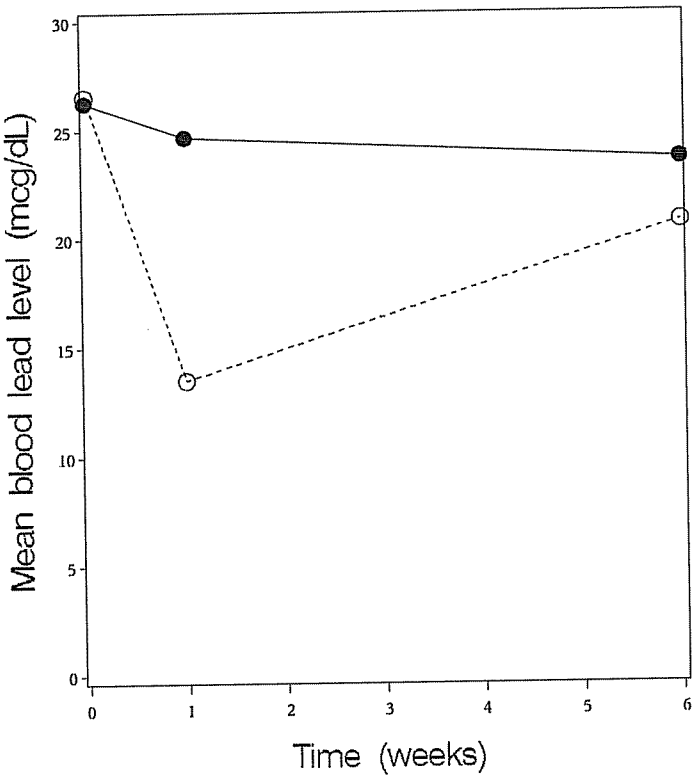


Figure 1: *Mean blood levels ($\mu g/dL$) per time point for children from the TLC trial. Black dots correspond to the placebo group and open circles correspond to the succimer group.*

Hypothesis B is more appropriate. If we fail to reject null hypothesis B we conclude that all changes in mean response from baseline could be the same in the two treatment groups. However, we do not also imply that mean responses at each time point could be equal across the two groups. Null hypothesis A, on the other hand, also requires that the mean responses at each time point be equal in the two treatment groups.

3. Consider only the succimer group and test whether there was a change in mean blood lead levels between baseline ($j = 1$) and week 1 ($j = 2$).

We formulate a simple hypothesis:

$$H_0 : \mu_{S2} - \mu_{S1} = 0, \qquad H_a : \mu_{S2} - \mu_{S1} \neq 0.$$

Let $\bar{Y}_{Sj}$ denote the mean response at the jth time point in the succimer group. The test statistic is:

$$t = \frac{\bar{Y}_{S2} - \bar{Y}_{S1}}{\text{SE}(\bar{Y}_{S2} - \bar{Y}_{S1})} = \frac{14 - 27}{\sqrt{50^{-1}(25.2 + 58.8 - 2 \times 15.4)}}$$

$$= \frac{-13}{1.04} = -12.5.$$

Since $|t| > 2$ which is approximately equal to the upper 2.5th percentile of a t-distribution with 49 degrees of freedom, we reject H_0.

4. An alternative design for a study with a similar objective consists in randomly allocating 150 children to each treatment group but then measuring a random sub-sample of 50 children in each treatment group at each measuring occasion. In this design, we also obtain a total of 300 measurements, but each child provides only one measurement of blood lead levels. Which design (this, or the design used in the *TLC* study) would you prefer and why?

   The longitudinal study, where children are measured more than once, is the preferred design as long as the correlation between measurements taken on the same child is positive. Intuitively, when computing the variance of the difference between two means on the same set of children, the variance will be smaller if the covariance between the two means is positive. For example, the SD of the difference $\bar{Y}_{S2} - \bar{Y}_{S1}$ increases to 1.30 (from 1.04) if a different and independent set of children had been measured at baseline and at week 1. In order to keep the SD of the mean difference approximately equal to 1.04, it would have been necessary to measure approximately 78 children per treatment group and per occasion which greatly increases the overall sample size to 468.

Part II

Refer to Figure 1 and Table 2. An approach to comparing the mean responses over time in the succimer and the placebo groups consists in computing the *areas under the curves* (*AUC*) between baseline ($t_1 = 0$) and week 6 ($t_3 = 6$). The shaded areas in Figure 2 show the sample *AUC* for the placebo and the succimer groups on the left and right panels, respectively.

5. Let AUC_P and AUC_S denote population versions of the sample areas displayed in the left and right panels of Figure 2, respectively. Conduct a test of the hypothesis that AUC_P equals AUC_S. Assume that the response vectors for all children in a treatment group are 3-dimensional multivariate normal random vectors.

   The area under each of the two group curves can be estimated as the sum of the areas of two trapezoids. For the placebo group, the AUC is:

$$\begin{aligned}
\widehat{AUC}_P &= (t_2 - t_1) \times \frac{\bar{Y}_{P1} + \bar{Y}_{P2}}{2} + (t_3 - t_2) \times \frac{\bar{Y}_{P2} + \bar{Y}_{P3}}{2} \\
&= \frac{1}{2}[(t_2 - t_1)\bar{Y}_{P1} + (t_3 - t_1)\bar{Y}_{P2} + (t_3 - t_2)\bar{Y}_{P3}] \\
&= \mathbf{L}'\bar{\mathbf{Y}}_P,
\end{aligned}$$

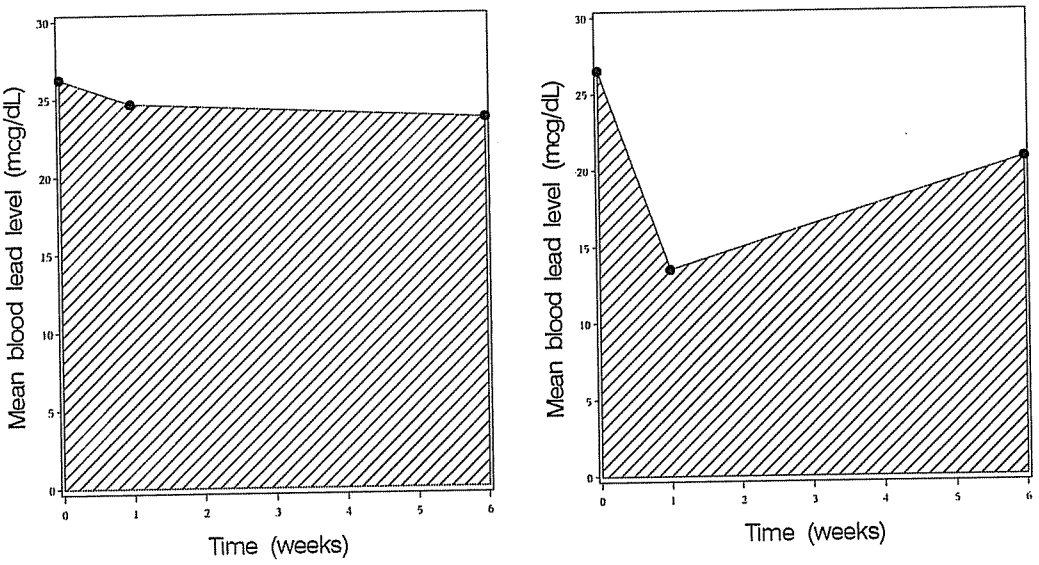


Figure 2: *Area under the response curves. The left panel corresponds to the placebo group and the right panel corresponds to the succimer group.*

with

$$\mathbf{L} = \frac{1}{2}\begin{bmatrix} t_2 - t_1 \\ t_3 - t_1 \\ t_3 - t_2 \end{bmatrix}, \quad \bar{\mathbf{Y}}_P = \begin{bmatrix} \bar{Y}_{P1} \\ \bar{Y}_{P2} \\ \bar{Y}_{P3} \end{bmatrix}.$$

To calculate $\widehat{AUC}_S$ for the succimer group we repeat the calculation above. The linear combination of measurement occasion means for the succimer group is denoted $\mathbf{L}'\bar{\mathbf{Y}}_S$.

Then the null hypothesis of no differences between the two areas under the response curves is:

$$H_0 : \mathbf{L}'\boldsymbol{\mu}_S = \mathbf{L}'\boldsymbol{\mu}_P \longrightarrow \mathbf{L}'(\boldsymbol{\mu}_S - \boldsymbol{\mu}_P) = 0.$$

There are different strategies to carry out the test of hypothesis under the assumption of normality.

The first strategy consists in using a one-degree of freedom Wald test, with test statistic

$$W^2 = \mathbf{L}'(\bar{\mathbf{Y}}_S - \bar{\mathbf{Y}}_P)(\mathbf{L}'\text{Cov}(\bar{\mathbf{Y}}_S - \bar{\mathbf{Y}}_P)\mathbf{L})^{-1}(\bar{\mathbf{Y}}_S - \bar{\mathbf{Y}}_P)'\mathbf{L},$$

which has an asymptotic χ^2 distribution with 1 df.

A second strategy is to carry out a two-sample t-test and use the approach proposed by Welch to approximate the degrees of freedom corresponding to the

linear combination of two means with different sample variances. Let

$$S_P^2 = \hat{\text{Var}}(\mathbf{L}'\mathbf{Y}_P), \quad S_S^2 = \hat{\text{Var}}(\mathbf{L}'\mathbf{Y}_S).$$

Then the test statistic is:

$$t = \frac{\mathbf{L}'\bar{\mathbf{Y}}_P - \mathbf{L}'\bar{\mathbf{Y}}_S}{\sqrt{\frac{S_P^2}{50} + \frac{S_S^2}{50}}},$$

and the corresponding degrees of freedom ν are given by:

$$\nu = \left(\frac{S_P^2}{50} + \frac{S_S^2}{50}\right)^2 \left(\frac{S_P^4}{50^2(50-1)} + \frac{S_S^4}{50^2(50-1)}\right)^{-1}.$$

The vector $\mathbf{L}'$ is equal to:

$$\mathbf{L}' = \frac{1}{2}\begin{bmatrix} 1 & 6 & 5 \end{bmatrix}.$$

The vector of mean response differences at each time point is:

$$\bar{\mathbf{Y}}_P - \bar{\mathbf{Y}}_S = \begin{bmatrix} -1 \\ 11 \\ 3 \end{bmatrix},$$

and the covariance matrix of the vector of mean response differences is:

$$\begin{aligned} \text{Cov}(\bar{\mathbf{Y}}_P - \bar{\mathbf{Y}}_S) &= \frac{1}{50}(\mathbf{V}_S + \mathbf{V}_P) \\ &= \frac{1}{50}\begin{bmatrix} 50.4 & 38.2 & 44.7 \\ 38.2 & 88.6 & 59.6 \\ 44.7 & 59.6 & 117.4 \end{bmatrix}. \end{aligned}$$

Then

$$\begin{aligned} \mathbf{L}'(\bar{\mathbf{Y}}_P - \bar{\mathbf{Y}}_S) &= \frac{1}{2}\begin{bmatrix} 1 & 6 & 5 \end{bmatrix}\begin{bmatrix} -1 \\ 11 \\ 3 \end{bmatrix} = \frac{80}{2} = 40, \\ \mathbf{L}'\text{Cov}(\bar{\mathbf{Y}}_P - \bar{\mathbf{Y}}_S)\mathbf{L} &= \frac{1}{4}\begin{bmatrix} 1 & 6 & 5 \end{bmatrix}\frac{1}{50}\begin{bmatrix} 50.4 & 38.2 & 44.7 \\ 38.2 & 88.6 & 59.6 \\ 44.7 & 59.6 & 117.4 \end{bmatrix}\begin{bmatrix} 1 \\ 6 \\ 5 \end{bmatrix} \\ &= \frac{1}{4} \times \frac{1}{50} \times 10,656.4 = 53.28. \end{aligned}$$

We first compute the Wald statistic:

$$W^2 = 40\,(44.3)^{-1}\,40 = 30.03.$$

We compare the test statistic to the $1-\alpha$ quantile of a χ^2 distribution with 1 d.f. and conclude that there is a significant difference between the two areas. The area under the response curve corresponding to the placebo group is significantly larger than the area under the response curve corresponding to the succimer group.

Alternatively, we carry out a two-sample t-test, where

$$\mathbf{L}'\bar{\mathbf{Y}}_P = 148, \quad \mathbf{L}'\bar{\mathbf{Y}}_S = 108, \quad S_P^2 = 946.2, \quad S_S^2 = 1717.7.$$

The t-statistic and its approximate degrees of freedom are

$$\begin{aligned} t &= (148-108)\left(\frac{946.2}{50}+\frac{1717.7}{50}\right)^{-1/2} \\ &= \frac{40}{7.3} = 5.48, \\ \nu &= \left(\frac{946.2}{50}+\frac{1717.7}{50}\right)^2\left(\frac{946.2^2}{50^2(49)}+\frac{1717.7^2}{50^2(49)}\right)^{-1} \\ &= \frac{2838.97}{31.4} = 90.4 \approx 90. \end{aligned}$$

For an $\alpha-$level of 0.05 we reject the null hypothesis of no differences if the absolute value of the statistic exceeds (approximately) the critical value 2. Therefore, we reject the null hypothesis and conclude that the areas under the two response curves are significantly different.

Part III

Two models were fit to these data. In both cases, the vector of blood lead levels for the kth subject in the ith treatment group, $\mathbf{Y}_{ik}$, was assumed to be $\mathrm{N}_3(\mathbf{X}_{ik}\boldsymbol{\beta}, \boldsymbol{\Sigma})$, with $\mathbf{X}_{ik}$ a $3 \times p$ design matrix, p the dimension of the vector of unknown regression coefficients $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ the 3×3 covariance matrix. The response vectors for the $N = 100$ children in the study are assumed to be independent. The covariance matrix $\boldsymbol{\Sigma}$ was assumed to be homogeneous across treatment groups. Treatment group is indicated by the dummy variable g_i that takes on values 0 or 1 for placebo and succimer groups respectively. In both models, the covariance matrix $\boldsymbol{\Sigma}$ has a compound symmetric structure, so that:

$$\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}_3 + \sigma_b^2\boldsymbol{J}_3,$$

where $\boldsymbol{I}_3$ is a 3×3 identity matrix, $\boldsymbol{J}_3$ is a 3×3 matrix of ones, and σ_b^2, σ^2 are unknown scalar-valued parameters with $\sigma_b^2 > 0$ and $\sigma^2 > 0$.

Model 1 was a linear time trend model with

$$E(Y_{ijk}) = \beta_0 + \beta_1 g_i + \beta_2 t_j + \beta_3 g_i t_j.$$

Model 2 was a linear spline model, with a knot at week 1, so that

$$\begin{aligned} E(Y_{ijk}) &= \beta_0 + \beta_1 g_i + \beta_2 t_j + \beta_3 g_i t_j \\ &\quad + \beta_4 (t_j - 1)_+ + \beta_5 g_i (t_j - 1)_+, \end{aligned}$$

where $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$.

6. Under Model 1 provide an expression for $E(Y_{ijk})$ for all combinations of $i \in \{P, S\}$ and $j = 1, 2, 3$.

	$i = P$	$i = S$
$j = 1$	β_0	$\beta_0 + \beta_1$
$j = 2$	$\beta_0 + \beta_2$	$(\beta_0 + \beta_1) + (\beta_2 + \beta_3)$
$j = 3$	$\beta_0 + \beta_2 6$	$(\beta_0 + \beta_1) + (\beta_2 + \beta_3)6$

7. Under Model 2 provide an expression for $E(Y_{ijk})$ for all combinations of $i \in \{P, S\}$ and $j = 1, 2, 3$.

	$i = P$	$i = S$
$j = 1$	β_0	$\beta_0 + \beta_1$
$j = 2$	$(\beta_0 - \beta_4) + (\beta_2 + \beta_4)$	$(\beta_0 + \beta_1 - \beta_4 - \beta_5) + (\beta_2 + \beta_3 + \beta_4 + \beta_5)$
$j = 3$	$(\beta_0 - \beta_4) + (\beta_2 + \beta_4)6$	$(\beta_0 + \beta_1 - \beta_4 - \beta_5) + (\beta_2 + \beta_3 + \beta_4 + \beta_5)6$

8. Use the attached SAS code and output to estimate the variance of the following differences under Model 1:

a) $Y_{ijk} - Y_{ijk'}$, $k \neq k'$

b) $Y_{ijk} - Y_{ij'k'}$, $j \neq j'$; $k \neq k'$

c) $Y_{ijk} - Y_{ij'k}$, $j \neq j'$.

a) and b) correspond to response differences between different children at respectively the same measurement occasion or at different measurement occasions. The third difference is of the responses of the same child at different

measurement occasions. The estimated variances of the three differences are:

$$\begin{aligned}
\widehat{Var}(Y_{ijk} - Y_{ijk'}) &= (\hat{\sigma}_b^2 + \hat{\sigma}^2) + (\hat{\sigma}_b^2 + \hat{\sigma}^2) = 2(\hat{\sigma}_b^2 + \hat{\sigma}^2) = 2(40.4 + 16.6) = 114.0, \\
\widehat{Var}(Y_{ijk} - Y_{ij'k'}) &= (\hat{\sigma}_b^2 + \hat{\sigma}^2) + (\hat{\sigma}_b^2 + \hat{\sigma}^2) = 2(\hat{\sigma}_b^2 + \hat{\sigma}^2) = 2(40.4 + 16.6) = 114.0, \\
\widehat{Var}(Y_{ijk} - Y_{ij'k}) &= (\hat{\sigma}_b^2 + \hat{\sigma}^2) + (\hat{\sigma}_b^2 + \hat{\sigma}^2) - 2\hat{\sigma}_b^2 = 2\hat{\sigma}_b^2 + 2\hat{\sigma}^2 - 2\hat{\sigma}_b^2 \\
&= 33.2 + 80.8 - 33.2 = 80.8.
\end{aligned}$$

The correlation between two measurements taken on the same child can be estimated as $\hat{\rho} = \hat{\sigma}_b^2/(\hat{\sigma}_b^2 + \hat{\sigma}^2)$.

9. Refer to the attached SAS code and output. Under Model 1, test the hypothesis that the changes in mean lead blood levels over time in the succimer group are the same as those in the placebo group.

Given the parametrization in terms of regression coefficients, in the linear model we can test for differences in the effect of treatment on change in the response by testing a hypothesis about β_3. If β_3 is not significantly different from zero, we conclude that the change in response is the same in the two groups of children.

The test is as follows:

$$H_0 : \beta_3 = 0 \quad H_a : \beta_3 \neq 0.$$

The test statistic is:

$$t = \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)}.$$

From the SAS output we find that

$$t = \frac{0.177}{0.280} = 0.63.$$

Comparing this statistic to a $t-$distribution with 198 d.f., we conclude that under Model 1, there is no evidence of a different effect of treatment on the changes in mean blood lead levels over time.

10. Refer to the attached SAS code and output. Under Model 2, test the hypothesis that the changes in mean lead blood levels over time in the succimer group are the same as those in the placebo group.

There are different approaches to test whether there are treatment effects on the change in mean response under Model 2. The most direct approach is to simultaneously test whether β_3 and β_5 are different from zero.

We first test the general null hypothesis

$$H_0 : \boldsymbol{C\beta} = 0,$$

against the alternative that $\mathbf{C}\boldsymbol{\beta}$ is not equal to zero. To test whether both β_3, β_5 are equal to zero, we let

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

A Wald statistic for this test is given by

$$W^2 = (\mathbf{C}\boldsymbol{\beta})'(\mathbf{C}\hat{\mathbf{V}}_B\mathbf{C}')^{-1}(\mathbf{C}\boldsymbol{\beta}),$$

where $\hat{\mathbf{V}}_B$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$. The test statistic W^2 is asymptotically distributed as a χ^2 random variable with degrees of freedom equal to q, the number of linearly independent rows in $\mathbf{C}$. Here, $q = 2$.

From the SAS output we get

$$\begin{aligned} C\hat{\beta} &= \begin{bmatrix} -11.41 \\ 13.06 \end{bmatrix}, \\ \mathbf{C}\hat{\mathbf{V}}_B\mathbf{C}' &= \begin{bmatrix} 1.526 & -1.679 \\ -1.679 & 1.893 \end{bmatrix}. \end{aligned}$$

Then, the W^2 statistic is equal to:

$$\begin{aligned} W^2 &= \begin{bmatrix} -11.41 & 13.06 \end{bmatrix} \begin{bmatrix} 1.526 & -1.679 \\ -1.679 & 1.893 \end{bmatrix}^{-1} \begin{bmatrix} -11.41 \\ 13.06 \end{bmatrix} \\ &= \begin{bmatrix} -11.41 & 13.06 \end{bmatrix} \begin{bmatrix} 27.168 & 24.097 \\ 24.097 & 21.901 \end{bmatrix} \begin{bmatrix} -11.41 \\ 13.06 \end{bmatrix} \\ &= 90.89. \end{aligned}$$

The 0.95 quantile of a χ^2_2 equals 5.99. Since $W^2 > 5.99$ we reject the null hypothesis and conclude that $\mathbf{C}\boldsymbol{\beta}$ is not equal to zero at the 5% level.

Other, less powerful approaches consist in testing whether there are group differences during the two periods on or before week 1 and after week 1, separately.

In the first period, we can test whether β_3 is significantly different from zero to decide whether the responses before week 1 are different for the succimer and the placebo groups. To do so, we set up the null hypothesis $H_0 : \beta_3 = 0$. The test statistic is:

$$t = \frac{\hat{\beta}_3}{SE(\hat{\beta}_3)},$$

where $\hat{\beta}_3 = -11.41$ with $SE = 1.25$, which is significant at the 5% level.

To test for treatment differences on changes in the response over time in the period after week 1, the null hypothesis that we need to test is $H_0 : \beta_3 + \beta_5 = 0$.

Recall that the slopes of the two treatments differ by $\beta_3 + \beta_5$ in the second period. The test statistic is:

$$t = \frac{\hat{\beta}_3 + \hat{\beta}_5}{SE(\hat{\beta}_3 + \hat{\beta}_5)}.$$

From the SAS output we find that

$$\begin{aligned} \hat{\beta}_3 + \hat{\beta}_5 &= 1.651, \\ \text{Var}(\hat{\beta}_3 + \hat{\beta}_5) &= \text{Var}(\hat{\beta}_3) + \text{Var}(\hat{\beta}_5) + 2\text{Cov}(\hat{\beta}_3, \hat{\beta}_5) \\ &= 1.526 + 1.893 + 2 \times (-1.679) = 0.061. \end{aligned}$$

Therefore, $t = 6.68$ which leads to rejection of H_0 at the 5% level. We conclude that the slopes of mean blood lead values on time after week 1 are also different in the two treatment groups.

The joint α-level for the tests of hypothesis in the two periods is more than 5%. The first approach, where both regression coefficients are tested simultaneously is preferred.

Corn stover consists of the leaves and stalks of maize plants left in a field after grain is harvested. As stover decays, it emits carbon dioxide (CO_2) into the atmosphere. The Intergovernmental Panel on Climate Change has suggested that mixing corn stover and other similar crop residues into soil after grain harvest may partially trap CO_2 in the soil and thereby reduce atmospheric levels of CO_2. For this reason, researchers are interested in studying variation among maize genotypes with respect to CO_2 emission. One goal of the research is to identify genotypes that are relatively slow to emit CO_2 when mixed with soil.

Suppose an experiment was conducted to study CO_2 emission of 8 maize genotypes. The 8 maize genotypes were planted on 16 plots in a field. The 16 plots were arranged in 2 blocks of 8 plots each, and a randomized complete block design was used to assign the 8 maize genotypes to the 16 plots. Following harvest, 2 samples of stover mixed with soil were collected from each of the 16 plots. Each sample of stover mixed with soil was placed in a canister. The 32 canisters were then randomly arranged in a laboratory where measurements of CO_2 emission were collected over a period of several days.

Unfortunately, it was not possible to continuously monitor CO_2 emission from each canister. Instead, CO_2 emission from each canister was measured for a brief period of time on each of several measurement days. These measurements were used to obtain the estimates of the amount of CO_2 emitted per gram of stover per day displayed in Table 1 on page 7. Throughout this problem, y_{ijkl} will denote the response measured on day l for the k^{th} canister of genotype j from block i ($i = 1, 2; j = 1, \ldots, 8; k = 1, 2; l = 1, 2, 4, 8, 16, 32, 64$).

Part I

Suppose the researchers would like to begin with an analysis of the day 1 data to determine if there are any significant differences among genotypes with respect to CO_2 emission on day 1. When answering questions in Part I, assume that only the day 1 data are available. Suppose for $i = 1, 2; j = 1, \ldots, 8;$ and $k = 1, 2;$

$$y_{ijk1} = \mu + \delta_i + \gamma_j + p_{ij} + e_{ijk}, \tag{1}$$

where $\mu, \delta_1, \delta_2, \gamma_1, \ldots, \gamma_8$ are unknown, real valued parameters; the p_{ij} terms are iid $N(0, \sigma_p^2)$; the e_{ijk} terms are iid $N(0, \sigma_e^2)$; and the p_{ij} terms are independent of the e_{ijk} terms.

1. In the model formulation above, σ_p^2 and σ_e^2 denote unknown, positive variance components. Describe in words what each of these variance components represents.

2. Model (1) can be written simultaneously for all data in the form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e},$$

where the order of elements in $\boldsymbol{y}$ matches the order of the responses in the 5^{th} column of Table 1 and

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R} \end{bmatrix}\right).$$

Provide expressions for $\boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{Z}, \boldsymbol{u}, \boldsymbol{G}$, and $\boldsymbol{R}$. You are encouraged to use Kronecker product notation to avoid writing out every element in large matrices. You are welcome to use $\boldsymbol{I}_{n\times n}$ to denote an identity matrix with n rows and n columns and $\boldsymbol{0}_{r\times c}$ and $\boldsymbol{1}_{r\times c}$ to denote matrices of zeros and ones, respectively, with r rows and c columns. If you do use this type of notation, please be sure to specify the dimensions.

3. Provide a simplified expression for $\text{var}(\boldsymbol{y})$. Kronecker product notation is recommended.

4. State the correlation between the day 1 responses from the two canisters corresponding to any single plot.

5. Provide a simplified expression for the best linear unbiased estimator of $\gamma_1 - \gamma_2$. Note that you are not being asked to compute an estimate using the data, and you are not required to derive the best linear unbiased estimator. Simply state a formula for the estimator in terms of y_{ijk1} $(i = 1, 2; j = 1, \ldots, 8; k = 1, 2)$.

6. Provide a simplified expression for the variance of the best linear unbiased estimator of $\gamma_1 - \gamma_2$.

7. Provide a simplified expression for an unbiased estimator of $\gamma_1 - \gamma_2$ that is not the best linear unbiased estimator.

8. Provide a simplified expression (in terms of the variance components) for the variance of the unbiased estimator of $\gamma_1 - \gamma_2$ provided in question 7.

9. Use the `R` code and partial output provided below to answer the following questions. Note that `block`, `geno`, and `plot` are factors in `R` whose entries correspond to the first three columns of Table 1. The numeric vector `y` corresponds to the 5^{th} column of Table 1.

```
> anova(lm(y~block+geno+plot))
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq  F value    Pr(>F)
block                14.7968 116.9072 9.168e-09 ***
geno                  6.9699  55.0683 5.164e-10 ***
plot                  0.3505   2.7689   0.04341 *
Residuals             0.1266
```

   a) State the degrees of freedom for each row of the ANOVA table.

   b) Assuming that model (1) holds, provide an appropriate F-statistic for testing

$$H_0 : \gamma_1 = \cdots = \gamma_8.$$

10. Recall that the canisters were randomly arranged in a laboratory where measurements of CO_2 emission were collected. For this question, suppose that only one canister (rather than two) was collected for each of the 16 plots. Furthermore, suppose 4 different laboratory technicians will divide up the work of measuring the CO_2 emissions from the 16 canisters.

a) Specify which canisters you would like each of the 4 technicians to measure. Assume that each canister can be measured only once. You may refer to each canister by an ordered pair that indicates the block from the field experiment (1 or 2) and genotype (1, 2, 3, 4, 5, 6, 7, or 8) associated with the canister. For example, (1,3) denotes the canister associated with the plot in block 1 that was assigned genotype 3. Your answer should be a list of ordered pairs for each of the 4 technicians.

b) Assume that an additive Gauss-Markov linear model that includes fixed effects for field blocks, technicians, and genotypes is appropriate for the day 1 data resulting from the design you specified in question 10(a). Provide a design matrix $\boldsymbol{X}$ that corresponds to your design.

c) Based on your recommended design in question 10(a) and the additive model described in 10(b), are all contrasts of genotype effects estimable? Explain the reasoning behind your answer, or if possible, prove that your answer is correct.

Part II

Now suppose that the researchers would like to analyze all the data in Table 1. To simplify modeling and analysis, we will average the data over the two canisters for each plot separately for each day to obtain

$$\bar{y}_{ij\cdot l} = \sum_{k=1}^{2} y_{ijkl}/2 \text{ for } i = 1, 2; j = 1, \ldots, 8; l = 1, 2, 4, 8, 16, 32, 64.$$

It might be useful to think of obtaining these data by averaging pairs of adjacent rows (1 with 2, 3 with 4, 5 with 6, etc.) in Table 1. We will use $\bar{\boldsymbol{y}}$ to denote the vector of these observations ordered first by block, then by genotype, and finally by day.

Consider a general linear model

$$\bar{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2}$$

where $\boldsymbol{X}$ is a design matrix, $\boldsymbol{\beta}$ is a vector of fixed parameters, and $\boldsymbol{e} \sim N(\boldsymbol{0}, \Sigma)$. The matrix Σ is assumed to be a positive definite matrix whose structure may be known but whose entries may depend on the unknown parameters. Three versions of this model, using three different choices for Σ, were fit using the R code that begins at the top of next page.

```
m1=gls(y~block+geno*day,data=da,
       correlation = corCompSymm(value=0,fixed=T,form=~1|plot),
       method="REML")

m2=gls(y~block+geno*day,data=da,
       correlation = corCompSymm(form=~1|plot),
       method="REML")

m3=gls(y ~ block+geno*day,data=da,
       correlation = corAR1(form=~1|plot),
       method="REML")
```

Note that `da` is a data frame that contains the response vector $\bar{\boldsymbol{y}}$ (denoted in the data frame as `y`) along with factors `block`, `geno`, and `day`, which indicate block, genotype, and measurement day, respectively. Each of the three models includes the assumption that

$$\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_{16\times 16} \otimes \boldsymbol{W}_{7\times 7},$$

where $\boldsymbol{W}_{7\times 7} = [w_{uv}]_{u=1,\ldots,7;v=1,\ldots,7}$ is a positive definite matrix. In `m1`, it is assumed that

$$w_{uv} = \begin{cases} 1 & \text{for } u = v, \\ 0 & \text{for } u \neq v. \end{cases}$$

In `m2`, it is assumed that

$$w_{uv} = \begin{cases} 1 & \text{for } u = v, \\ \rho & \text{for } u \neq v, \end{cases}$$

where ρ is an unknown parameter in $(-1, 1)$. In `m3`, it is assumed that

$$w_{uv} = \begin{cases} 1 & \text{for } u = v, \\ \theta^{|u-v|} & \text{for } u \neq v, \end{cases}$$

where θ is an unknown parameter in $(-1, 1)$.

11. Determine the number of rows, number of columns, and rank of the design matrix $\boldsymbol{X}$ specified by the `R` code used to generate `m1`, `m2`, and `m3`.

12. Suppose that model (2) holds with design matrix $\boldsymbol{X}$ as specified by the `R` code used to generate `m1`, `m2`, and `m3`. Suppose $\boldsymbol{\Sigma}$ is a positive definite matrix whose entries depend on unknown variance parameters. For example, $\boldsymbol{\Sigma}$ could be one of the matrices discussed above, but the exact form of $\boldsymbol{\Sigma}$ is not specified in this problem. Recall that REML estimates of variance parameters are maximum likelihood estimates of variance parameters obtained by using a special linear transformation of the response vector as data. Suppose that $\boldsymbol{A}$ is a matrix such that the REML estimates of variance parameters maximize the likelihood of $\boldsymbol{A}'\bar{\boldsymbol{y}}$.

a) State the distribution of $A'\bar{y}$.

b) How many elements are in the vector $A'\bar{y}$?

c) What is the rank of the matrix A?

13. Consider the R command and partial output below.

```
> anova(m1,m2,m3)

   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m1     1                        -48.44
m2     2                        -28.94
m3     3                        -29.63
```

a) Consider comparing each pair of models using a likelihood ratio test. For each pair, either explain why a likelihood ratio test is not appropriate or conduct a likelihood ratio test. For any likelihood ratio test that is appropriate, provide the test statistic, its degrees of freedom, an approximate p-value, and an interpretation of the results.

b) Fill in the missing entries under AIC and BIC and explain which model for Σ is preferred based on each of these criterion.

14. The R results `m1`, `m2`, and `m3` each contain estimates of a parameter denoted by R as `geno2`.

a) Explain how to interpret this parameter.

b) Define the profile likelihood for the parameter `geno2`.

c) Explain how the profile likelihood for `geno2` could be used to obtain a confidence interval for `geno2`.

15. The researchers would like to obtain estimates of total CO_2 emission over the course of the study for each plot. To obtain such estimates, they will first fit the nonlinear function

$$\mathrm{E}(\mathrm{CO_2\ Emission}) = \alpha_1 + \alpha_2 \mathrm{e}^{-\alpha_3 \cdot \mathrm{day}}, \tag{3}$$

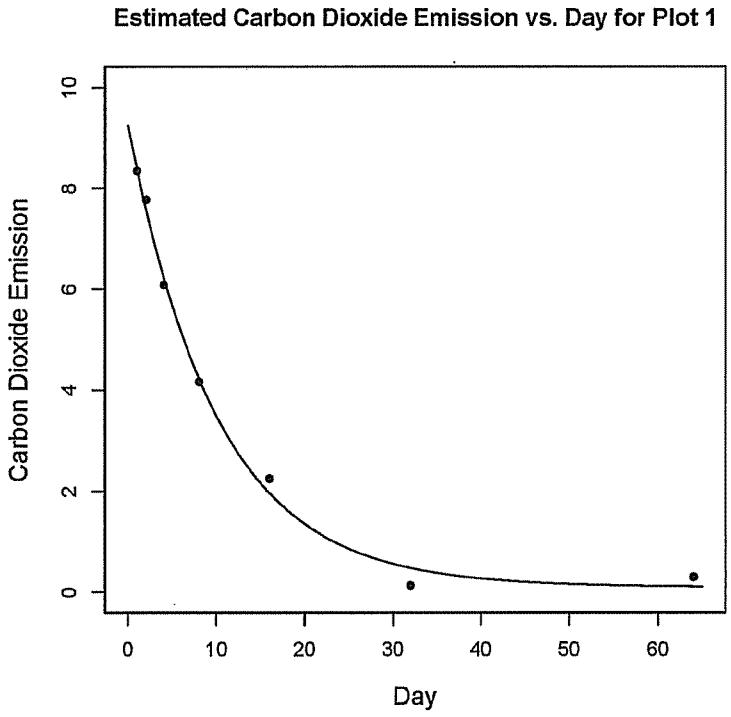
separately to the emissions data for each plot. Next, they will compute

$$z_{ij} \equiv \int_0^{64} \left(\hat{\alpha}_1^{(ij)} + \hat{\alpha}_2^{(ij)} e^{-\hat{\alpha}_3^{(ij)} \cdot x} \right) dx, \quad (i = 1, 2; j = 1, \ldots, 8)$$

where $\hat{\alpha}_1^{(ij)}$, $\hat{\alpha}_2^{(ij)}$, and $\hat{\alpha}_3^{(ij)}$ denote the least squares estimates of α_1, α_2, and α_3, respectively, for the plot from block i associated with genotype j. The quantity z_{ij} is the area under the estimated CO_2 emission curve and serves as an estimate of the total CO_2 emission over the course of the study for the plot from block i associated with genotype j.

a) Figure 1 depicts the least squares fit of function (3) to the data from the plot in block 1 associated with genotype 1. To fit the function (3) to data, it is necessary to provide starting values for the estimates of α_1, α_2, and α_3. Based on the data in Figure 1, suggest reasonable starting values.

Figure 1



b) The researchers are concerned that some estimates of total CO_2 emission may be more variable than others. Suppose $\hat{V}_{ij}$ is an estimate of $\text{var}([\hat{\alpha}_1^{(ij)}, \hat{\alpha}_2^{(ij)}, \hat{\alpha}_3^{(ij)}]')$. Provide an estimate of $\text{var}(z_{ij})$ as a function of $\hat{V}_{ij}$, $\hat{\alpha}_1^{(ij)}$, $\hat{\alpha}_2^{(ij)}$, and $\hat{\alpha}_3^{(ij)}$.

c) Let v_{ij} denote the variance estimate from question 15(b). The researchers would like to test for differences among genotypes using the z_{ij} measures of CO_2 emission as the response. They wish to use a weighted analysis that accounts for heterogeneity of variance among the z_{ij} estimates. What weight would you recommend assigning to z_{ij} in the weighted analysis?

d) Please briefly describe advantages and disadvantages of a weighted analysis as compared to an unweighted analysis.

Table 1. CO_2 Emissions Data

				Measurement Day						
Block	Genotype	Plot	Canister	1	2	4	8	16	32	64
1	1	1	1	8.57	8.13	6.25	4.38	2.40	0.25	0.43
1	1	1	2	8.12	7.40	5.92	3.96	2.11	0.01	0.17
1	2	2	3	6.28	5.05	2.88	1.55	0.52	0.64	0.20
1	2	2	4	6.72	5.16	3.18	1.70	0.61	0.58	0.81
1	3	3	5	8.89	7.10	4.71	2.71	1.78	0.81	0.65
1	3	3	6	8.88	8.04	5.43	2.80	1.51	1.6	0.68
1	4	4	7	8.71	7.68	6.39	4.50	2.44	1.14	1.30
1	4	4	8	9.23	8.38	7.08	5.20	2.95	0.49	0.76
1	5	5	9	9.04	8.16	6.68	4.30	1.89	0.98	0.30
1	5	5	10	9.01	8.04	6.76	5.07	2.2	1.38	0.23
1	6	6	11	8.88	8.23	6.94	4.91	1.92	1.46	0.59
1	6	6	12	8.27	7.36	6.12	4.56	1.80	1.04	0.52
1	7	7	13	5.00	3.38	1.66	0.79	1.10	0.4	1.03
1	7	7	14	5.24	3.70	1.55	0.66	1.01	1.37	1.04
1	8	8	15	9.09	8.09	6.76	4.42	2.41	0.92	0.53
1	8	8	16	8.38	7.52	5.95	4.27	1.90	0.63	0.48
2	1	9	17	10.22	9.29	8.06	6.13	3.92	2.27	1.58
2	1	9	18	10.01	9.15	7.77	5.78	3.19	1.99	2.28
2	2	10	19	8.13	6.65	4.32	2.49	1.11	1.80	1.81
2	2	10	20	8.08	6.19	4.30	3.00	2.21	2.46	2.53
2	3	11	21	9.37	7.92	5.89	3.70	2.11	2.48	1.36
2	3	11	22	10.38	9.07	6.80	4.51	2.82	1.94	2.14
2	4	12	23	9.22	8.33	7.10	5.13	2.92	1.21	1.20
2	4	12	24	9.47	8.37	7.14	5.12	3.01	1.31	0.47
2	5	13	25	9.39	8.78	8.03	5.99	4.77	2.71	1.89
2	5	13	26	10.36	9.47	8.34	6.81	4.48	3.08	2.67
2	6	14	27	10.48	9.55	7.36	5.99	3.88	2.17	2.09
2	6	14	28	11.06	10.11	8.88	6.98	4.15	2.62	3.31
2	7	15	29	6.79	5.15	3.35	2.20	2.29	2.37	2.33
2	7	15	30	6.97	5.60	4.11	2.42	1.97	2.03	1.95
2	8	16	31	10.12	8.83	7.79	5.46	3.44	1.64	1.59
2	8	16	32	10.02	9.28	7.49	5.65	4.16	2.46	1.45

1. The variance component σ_p^2 represents response variation among plots after accounting for the additive effects of blocks and genotypes. The variance component σ_e^2 represents response variation among canisters within a plot.

2.
$$\boldsymbol{X} = [\mathbf{1}_{32\times1}, \boldsymbol{I}_{2\times2} \otimes \mathbf{1}_{16\times1}, \mathbf{1}_{2\times1} \otimes \boldsymbol{I}_{8\times8} \otimes \mathbf{1}_{2\times1}]$$
$$\boldsymbol{\beta} = [\mu, \delta_1, \delta_2, \gamma_1, \ldots, \gamma_8]'$$
$$\boldsymbol{Z} = \boldsymbol{I}_{16\times16} \otimes \mathbf{1}_{2\times1}$$
$$\boldsymbol{u} = [p_{11}, p_{12}, \ldots, p_{18}, p_{21}, p_{22}, \ldots, p_{28}]'$$
$$\boldsymbol{G} = \sigma_p^2 \boldsymbol{I}_{16\times16}$$
$$\boldsymbol{R} = \sigma_e^2 \boldsymbol{I}_{32\times32}$$

3.
$$\boldsymbol{I}_{16\times16} \otimes \begin{bmatrix} \sigma_p^2 + \sigma_e^2 & \sigma_p^2 \\ \sigma_p^2 & \sigma_p^2 + \sigma_e^2 \end{bmatrix}$$

4. $\sigma_p^2/(\sigma_p^2 + \sigma_e^2)$

5. Let
$$\bar{y}_{\cdot j\cdot 1} = \sum_{i=1}^{2}\sum_{k=1}^{2} y_{ijk1}/4 \text{ for } j = 1, \ldots, 8.$$
Then $\bar{y}_{\cdot 1\cdot 1} - \bar{y}_{\cdot 2\cdot 1}$ is the best linear unbiased estimator.

6. $\text{var}(\bar{y}_{\cdot 1\cdot 1} - \bar{y}_{\cdot 2\cdot 1}) = \text{var}(\bar{p}_{\cdot 1} - \bar{p}_{\cdot 2} + \bar{e}_{\cdot 1\cdot 1} - \bar{e}_{\cdot 2\cdot 1}) = \sigma_p^2 + \sigma_e^2/2$

7. $y_{1111} - y_{1211}$

8. $\text{var}(y_{1111} - y_{1211}) = \text{var}(p_{11} - p_{12} + e_{1111} - e_{1211}) = 2\sigma_p^2 + 2\sigma_e^2$

9. **a)** 1,7,7,16 (in order from top to bottom)

   b) The plots are the experimental units, so an appropriate F-statistic is $6.9699/0.3505 \approx 19.89$.

10. **a)** Multiple answers are possible. It is important to nest the technician blocks inside the field blocks to allow for the most precise comparison of genotypes. It is also important to use a grouping of genotypes within technicians for the field block 1 data that is different from the grouping of genotypes within technicians for the field block 2 data. One reasonable design is as follows.

Technician 1	$(1,1)(1,2)(1,3)(1,4)$
Technician 2	$(1,5)(1,6)(1,7)(1,8)$
Technician 3	$(2,1)(2,2)(2,5)(2,6)$
Technician 4	$(2,3)(2,4)(2,7)(2,8)$

b) Because the technician blocks are nested within the field blocks and both blocking factors are considered fixed, the field blocks become irrelevant in the sense that the design matrix has the same column space with and without columns indicating the field blocks. This is easy to see because each column indicating a field block is the sum of columns for a pair of technicians. Thus, the following design matrix is appropriate for the design described in the answer to 10(a).

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{I}_{2\times 2} \otimes \mathbf{1}_{4\times 1} & \mathbf{0}_{8\times 2} & \boldsymbol{I}_{8\times 8} \\ \mathbf{0}_{8\times 2} & \boldsymbol{I}_{2\times 2} \otimes \mathbf{1}_{4\times 1} & \boldsymbol{A} \end{bmatrix},$$

where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

c) The mean of our response vector can be written as $\boldsymbol{X}\boldsymbol{\beta}$, where $\boldsymbol{\beta} = [\tau_1, \ldots, \tau_4, \gamma_1, \ldots, \gamma_8]'$. We wish to know whether $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable whenever $c_1 = c_2 = c_3 = c_4 = 0$ and $\sum_{i=5}^{12} c_i = 0$. A result from 511 states that $\boldsymbol{c}'\boldsymbol{\beta}$ is estimable if $\boldsymbol{c}'\boldsymbol{a} = 0$ whenever $\boldsymbol{X}\boldsymbol{a} = \mathbf{0}$. Suppose $\boldsymbol{X}\boldsymbol{a} = \mathbf{0}$ and that $\boldsymbol{c}$ is such that $c_1 = c_2 = c_3 = c_4 = 0$ and $\sum_{i=5}^{12} c_i = 0$. Then

$$\begin{aligned} a_1 + a_i &= 0 \quad \forall \quad i = 5, 6, 7, 8 \\ a_2 + a_i &= 0 \quad \forall \quad i = 9, 10, 11, 12 \\ a_3 + a_i &= 0 \quad \forall \quad i = 5, 6, 9, 10 \\ a_4 + a_i &= 0 \quad \forall \quad i = 7, 8, 11, 12 \end{aligned}$$

which implies

$$a_1 = a_2 = a_3 = a_4 = -d \text{ and } a_5 = \cdots = a_{12} = d$$

for some real number d. Thus,

$$\boldsymbol{c}'\boldsymbol{a} = \sum_{i=1}^{12} c_i a_i = \sum_{i=5}^{12} c_i d = d \sum_{i=5}^{12} c_i = 0.$$

It follows that all contrasts of genotype effects are estimable.

11. The design matrix has $16 \times 7 = 112$ rows. R uses set-first-to-zero constraints. Thus, there is 1 column for the intercept, 1 column for blocks, 7 columns for genotypes, 6 columns for days, and 42 columns for genotype by day interactions. This yields a total of 57 columns. The rank is also 57.

12. **a)** $\boldsymbol{A}$ is a matrix of linearly independent error contrasts. Thus, $\boldsymbol{A}'\boldsymbol{X} = \boldsymbol{0}$. It follows that

$$\boldsymbol{A}'\bar{\boldsymbol{y}} \sim N(\boldsymbol{0}, \boldsymbol{A}'\boldsymbol{\Sigma}\boldsymbol{A}).$$

b) The vector $\boldsymbol{A}'\bar{\boldsymbol{y}}$ should have $112 - 57 = 55$ elements because there are 55 residual degrees of freedom.

c) The matrix $\boldsymbol{A}$ is a matrix of linearly independent error contrasts. Thus, the rank must equal the number of columns, which is 55.

13. **a)** Model `m1` is a special case of model `m2` ($\rho = 0$) and a special case of `m3` ($\theta = 0$). Thus, we can test $H_0 : \rho = 0$ with the likelihood ratio statistic

$$-2(-48.44 + 28.94) = 39$$

and $H_0 : \theta = 0$ with the likelihood ratio statistic

$$-2(-48.44 + 29.63) = 37.62.$$

Based on a single degree of freedom, both null hypotheses are soundly rejected. Models `m2` and `m3` cannot be compared with a likelihood ratio test because neither one of the models is a special case of the other.

b)

$$AIC = -2\log\text{likelihood} + 2 * (\text{number of parameters})$$

$$BIC = -2\log\text{likelihood} + (\text{number of parameters}) * \log(\text{sample size})$$

Because the REML method was specified, R reports results for the REML likelihoods. Thus, the sample size for calculation of BIC is $112 - 57 = 55$. The number of parameters is considered to be 58, 59, and 59 for models 1, 2, and 3, respectively. This includes 57 for the mean parameters and 1, 2, and 2 for the variance parameters in models 1, 2, and 3, respectively. Thus, the resulting AIC and BIC values are as follows.

```
> aic=-2*anova(m1,m2,m3)$logLik+2*c(58,59,59)
> aic
[1] 212.8863 175.8889 177.2686
> bic=-2*anova(m1,m2,m3)$logLik+log(55)*c(58,59,59)
> bic
[1] 329.3116 294.3216 295.7013
> anova(m1,m2,m3)
   Model df      AIC      BIC    logLik   Test  L.Ratio p-value
m1     1 58 212.8863 329.3116 -48.44315
m2     2 59 175.8889 294.3216 -28.94447 1 vs 2 38.99736  <.0001
m3     3 59 177.2687 295.7013 -29.63432
```

It may be argued that the number of parameters should be only 1, 2, and 2, considering that the REML likelihood involves only the variance parameters. Using these values as the number of parameters is also acceptable and would not change the differences among AIC values or differences among BIC values. The differences among the values are more relevant than the absolute values themselves. Adding a constant to each AIC or to each BIC value has no effect on which model is preferred. In this case, the second model with the compound symmetric structure is preferred by both AIC and BIC.

14. **a)** This parameter represents the within-block difference between the mean for genotype 2 at day 1 and the mean for genotype 1 at day 1. This difference in means is the same for both block 1 and block 2 according to the model.

b) Let θ denote the vector of all model parameters including both fixed effects and variance components. Suppose θ_1, the first component of $\boldsymbol{\theta}$, corresponds to `geno2`. Let Ω denote the parameter space. Let $\Omega(\gamma_2) = \{\boldsymbol{\theta} \in \Omega : \theta_1 = \gamma_2\}$. Let $L(\boldsymbol{\theta})$ denote the likelihood function. The profile likelihood for `geno2` is defined for any real value γ_2 by

$$L^*(\gamma_2) = \sup\{L(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Omega(\gamma_2)\}.$$

c) The following is an approximate $100(1-\alpha)\%$ confidence interval for `geno2`:

$$\{\gamma_2 : \sup_{\boldsymbol{\theta}\in\Omega} \log[L(\boldsymbol{\theta})] - \log[L^*(\gamma_2)] \leq \chi^2_{1-\alpha}/2\},$$

where $\chi^2_{1-\alpha}$ is the $1-\alpha$ quantile of the chi-square distribution with 1 degree of freedom.

15. **a)** At day $= 0$, 10, and ∞, the function is equal to $\alpha_1 + \alpha_2$, $\alpha_1 + \alpha_2 \exp(-10\alpha_3)$, and α_1, respectively. Based on the observed data, values of the fitted function at day $= 0$, 10, and ∞ should be around 9, 4, and 0, respectively. Solving the system of three equations suggests starting values of 0 for $\hat{\alpha}_1$, 9 for $\hat{\alpha}_2$, and $-0.1 * \log(4/9) \approx 0.08$ for $\hat{\alpha}_3$.

b) To simplify notation, we will drop the (ij) superscripts. Carrying out the integration in the definition of z yields

$$z = 64\hat{\alpha}_1 + \hat{\alpha}_2 \left(1 - e^{-\hat{\alpha}_3 \cdot 64}\right) / \hat{\alpha}_3.$$

This is a nonlinear function of the parameters estimates. We can use the Delta method to obtain an estimated variance. Let

$$f(\alpha_1, \alpha_2, \alpha_3) = 64\alpha_1 + \alpha_2 \left(1 - e^{-\alpha_3 \cdot 64}\right) / \alpha_3$$

and

$$\boldsymbol{d} = [\partial f/\partial\alpha_1, \partial f/\partial\alpha_2, \partial f/\partial\alpha_3]'.$$

Let $\hat{d}$ denote $\boldsymbol{d}$ evaluated at the least squares estimates of α_1, α_2, and α_3. Straightforward differentiation shows that

$$\boldsymbol{d} = [64, \left(1 - e^{-\alpha_3 64}\right) / \alpha_3, \left(-\alpha_2 + 64\alpha_3 e^{-\alpha_3 64} + e^{-\alpha_3 64}\right) / \alpha_3^2]'.$$

An estimate of the variance of z is given by $\hat{\boldsymbol{d}}' \hat{\boldsymbol{V}}_{ij} \hat{\boldsymbol{d}}$.

c) The best linear unbiased estimator of contrasts of genotype effects weights each observation by the inverse of its variance. Thus, $1/v_{ij}$ may be the best available weight.

d) If the variance of each response were known, best linear unbiased estimation of genotype contrasts could be achieved using the weighted analysis with weights equal to the inverse of response variances. Ignoring the heterogeneity of variance could lead to inefficient estimation and potentially invalid inference. However, the variances are unknown and only estimated. If these estimates of the variances are poor and the actual heterogeneity of the variance is not severe, an equal-weights analysis may be preferred because the standard weighted analysis does not account for uncertainty in the weights.

Problem Background

Roughly 80% of the fruits and vegetables grown in the United States come from the state of California. Harvesting these crops, particularly vegetables, is labor intensive, and this creates a huge market for temporary agricultural workers. The amount of land devoted to the production of vegetables and fruits in California is an underlying factor in this demand. The National Agricultural Statistics Service (NASS) in the U.S. Department of Agriculture has compiled a historical record of the number of harvested acres in California considered to be in the categories of Vegetables and Melons, Fruits and Nuts, and Field Crops, which we will simply call "vegetables," "fruits," and "other." The data consist of records from 1960 through 2008 on the numbers of acres harvested in each of the three categories. We will use units of 100,000 acres.

The primary objective is to model patterns in these data over time, although forecasting future values would be of interest if it is possible. Plots of acres harvested in the three crop categories and the total of the three are presented in Figure 1 through Figure 4.

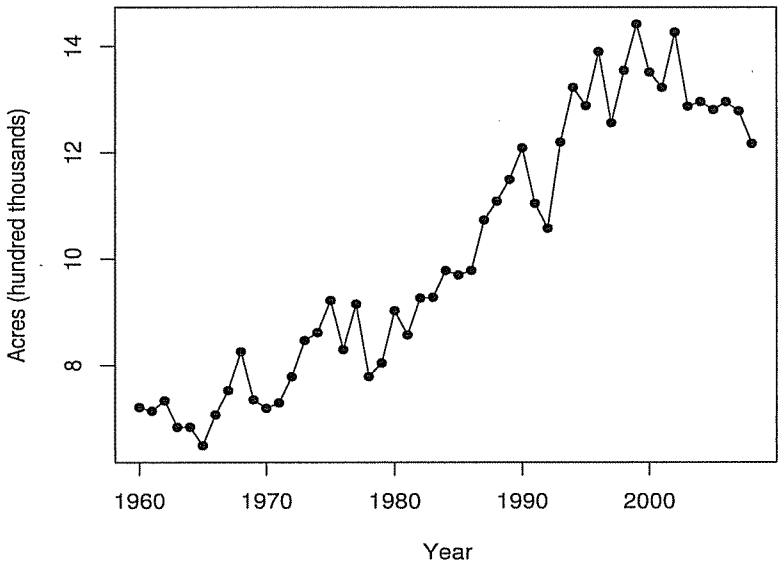


Figure 1: Acres in vegetables over time.

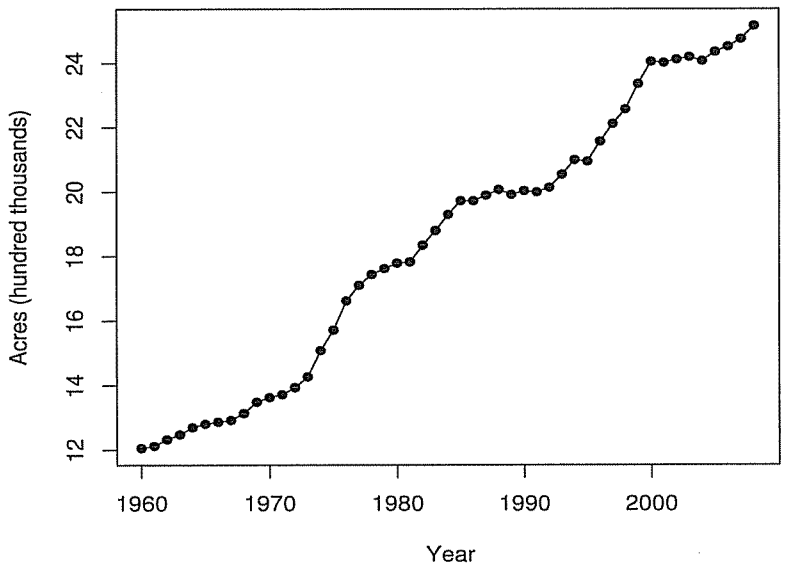


Figure 2: Acres in fruit over time.

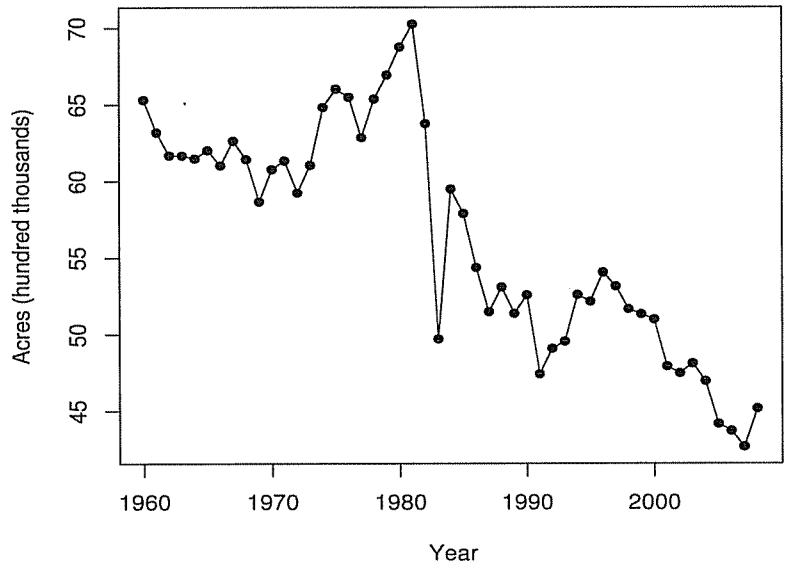


Figure 3: Acres in other crops over time.

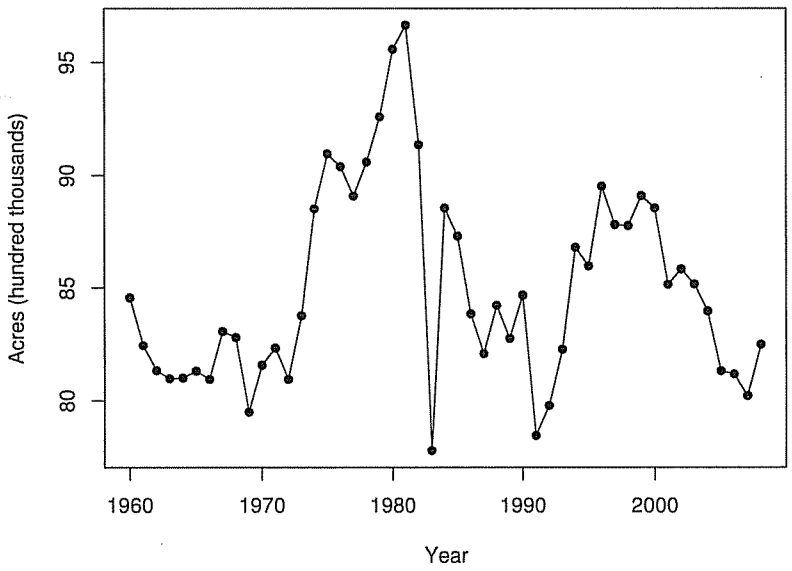


Figure 4: Acres in all crops over time.

Part I: Modeling Vegetable Acres

Attention is given first to modeling acres of vegetables over time. Let $Y(t)$ be a random variable connected with the number of harvested acres of vegetables in year $t = 1960, \ldots, 2008$, and let z_t be a time index such that $z_t = t - 1959$ for $t = 1960, \ldots, 2008$.

A simple linear regression model $Y(t) = \beta_0 + \beta_1 z_t + \sigma\epsilon_t$ with ϵ_t independent and identically distributed such that $E(\epsilon_t) = 0$ and $var(\epsilon_t) = 1$ was fit to the data of Figure 1. Estimation of regression parameters was by ordinary least squares, and the usual moment-based estimator was used for σ^2. Point and 95% interval estimates are presented in Table 1, and the coefficient of determination was $r^2 = 0.8869$. The fitted regression line and a plot of residuals against time is presented in Figure 5.

Parameter	Estimate	95% Interval
β_0	6.0497	(5.5579, 6.5415)
β_1	0.1634	(0.1463, 0.1805)
σ^2	0.7100	not produced

Table 1: Parameter estimates for fit of simple linear regression to the data of Figure 1.

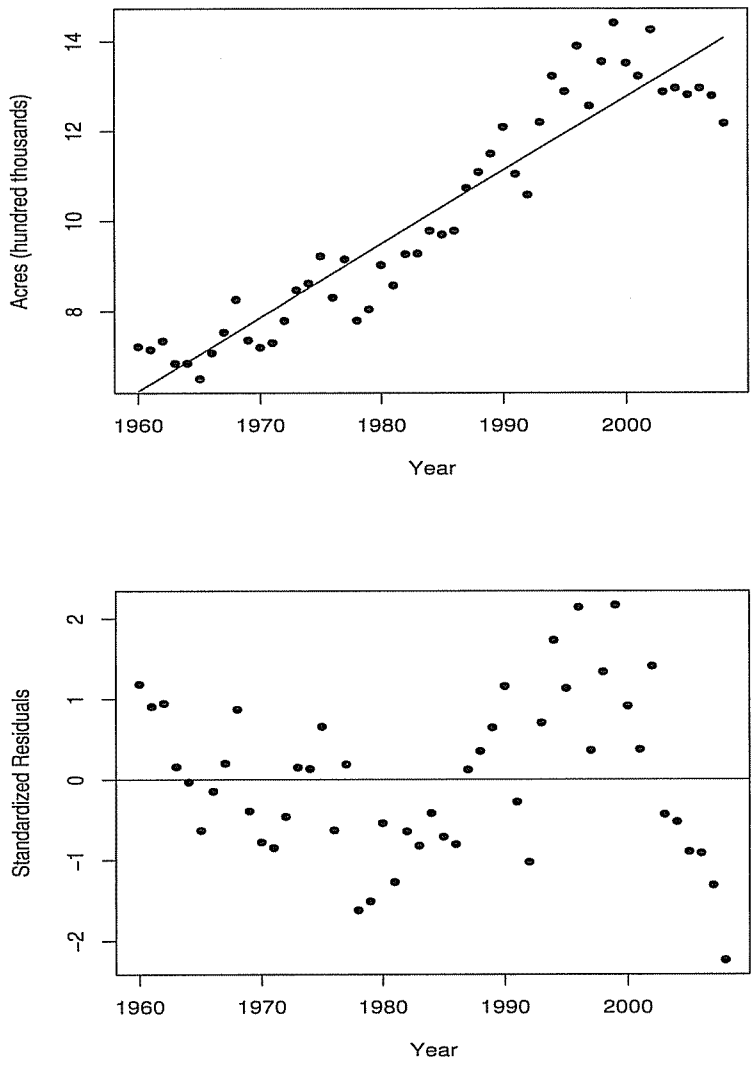


Figure 5: Fitted line and residuals from simple linear regression fit to vegetable data.

The residuals of Figure 5 appear to exhibit a pattern indicative of possible serial correlation over time. Exploratory tools (autocorrelation and partial autocorrelation plots, not shown) verify that the residuals exhibit the behavior of a first-order autoregressive process. Consider the following model as a possible improvement over the simple linear regression,

$$Y(t) = \beta_0 + \beta_1 z_t + X(t) \quad (1)$$

$$X(t) = \gamma X(t-1) + \sigma\,\epsilon_t, \quad (2)$$

where $\epsilon_t \overset{iid}{\sim} N(0,1)$ and $X(0) = 0$.

ANSWER QUESTIONS 1, 2, and 3 NOW. (SEE PAGE 14.)

Maximum likelihood estimates and 95% confidence intervals are shown in Table 2 for the parameters of the model specified by expressions (1) and (2).

Parameter	Estimate	95% Interval
β_0	6.2951	(5.3688, 7.2224)
β_1	0.1524	(0.1210, 0.1844)
σ^2	0.4196	(0.2429, 0.5700)
γ	0.6312	(0.2418, 0.7353)

Table 2: Maximum likelihood estimates and confidence intervals for a regression with autoregressive errors fit to the vegetable data of Figure 1.

Consider forecasting the number of vegetable acres over the next 20 years based on the model with estimates as given in Table 2. One way to accomplish this is to take the last marginal residual, $y(n) - \hat{\beta}_0 - \hat{\beta}_1 z_n$ (where $n = 2008$) and use it as a surrogate for the value of the autoregressive error process, $x(n)$. We might then simulate 20 further values from that process, $x(n+1), \ldots, x(n+20)$, using the estimated values $\hat{\gamma}$ and $\hat{\sigma}^2$ from Table 2 in the model of expression (2). Forecasts could be computed as $y(n+k) = \hat{\beta}_0 + \hat{\beta}_1 z_{n+k} + x(n+k)$ for $k = 1, \ldots, 20$. This whole process was repeated 5000 times. The mean value at each future time point was taken as the forecasted value at that time, and the 0.025 and 0.975 quantiles were taken as endpoints of a 95% forecast interval. These forecasts and forecast intervals are presented in Figure 6.

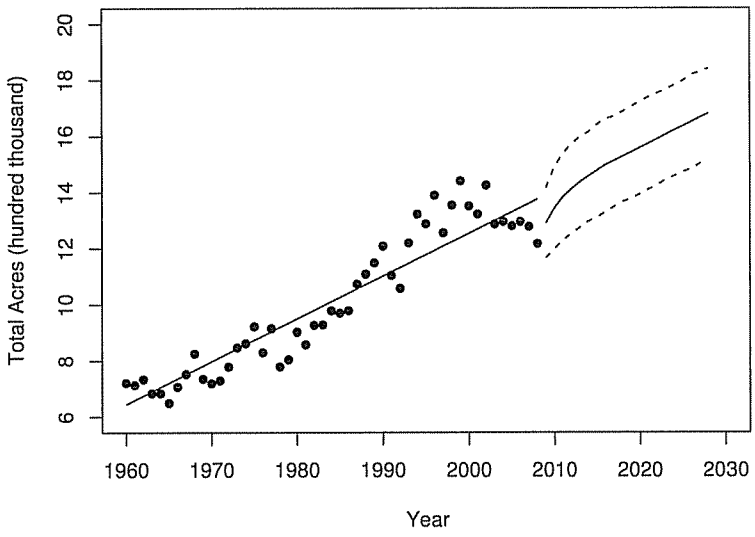


Figure 6: Forecasted vegetable acres harvested for next 20 years.

The forecasts of Figure 6 seem to revert back, in only a few years beyond 2008, to what would be given simply by the estimated expected values. To examine this further, a more detailed plot of forecasts and intervals for the years 2009-2014 from both the model with autoregressive errors and the original simple linear regression model with indpendent errors is presented in Figure 7.

ANSWER QUESTIONS 4, 5, 6 and 7 NOW. (SEE PAGE 15.)

Part II: Modeling Allocation of Agricultural Acres

Individual regression models for number of harvested acres in the three crop categories of vegetables, fruits, and other (of which we have considered only vegetables) perhaps do not provide the clearest picture of overall shifts in agricultural production in California. Rather, we might consider the proportions of total agricultural acres that are devoted to vegetables, fruits, and other crops. Plots of these proportions over time are presented in Figure 8, and an enlargement showing only vegetables and fruits is presented in Figure 9.

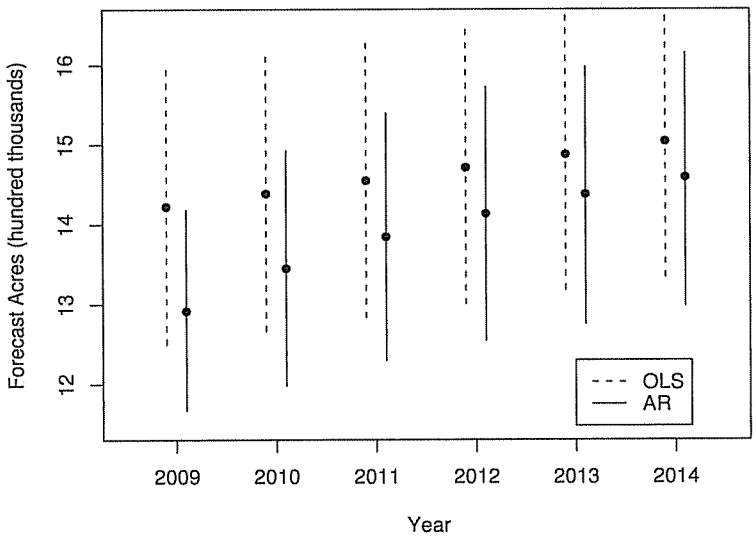


Figure 7: Forecasted vegetable acres from two models for 2009-2014.

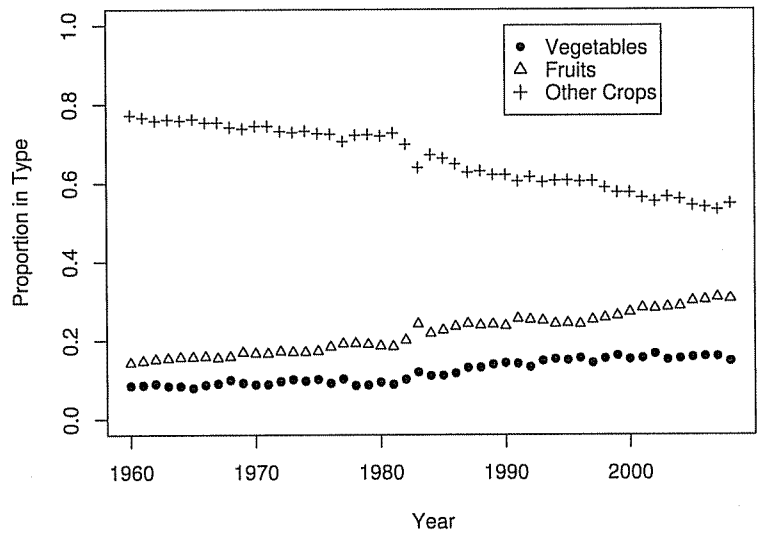


Figure 8: Proportions of total acres in crop categories over time.

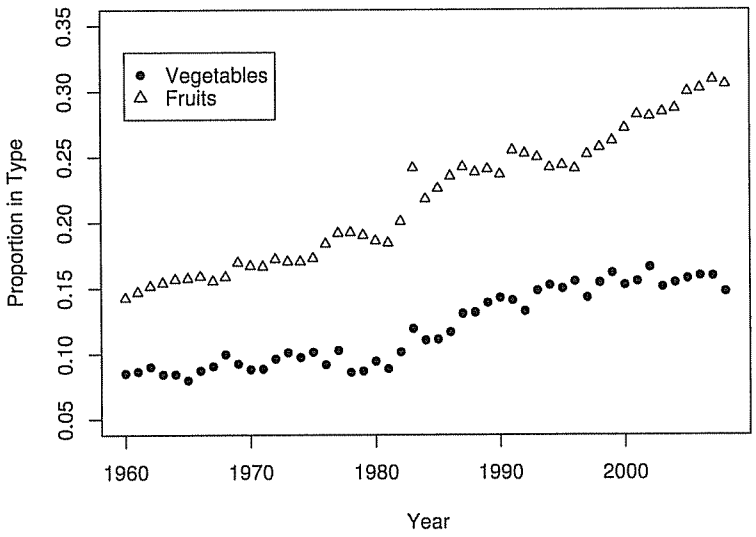


Figure 9: Proportions of total acres in vegetables and fruits over time.

We have interest in constructing a regression model that reflects changes over time in the proportional allocations of total acres to vegetables, fruits, and other crops. The literature contains a recent paper,

> Hijazi, R.H. and Jernigan, R.W. (2009), Modeling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics* **4**, 77-91.

In this paper the authors model proportions that sum to 1 using a Dirichlet distribution. In the context of the problem under consideration here, let $Y_v(t)$, $Y_f(t)$ and $Y_o(t)$ be random variables connected with the proportions of total agricultural acres devoted to vegetables, fruits, and other crops in year t, respectively. Note that for any possible values of these variables $y_v(t) + y_f(t) + y_o(t) = 1.0$ so that the value of any one is exactly determined by the values of the other two. If these random variables have a Dirichlet distribution then the joint probability density function may be written as, for $\alpha_{v,t} > 0$, $\alpha_{f,t} > 0$, $\alpha_{o,t} > 0$, $0 < y_v(t) < 1$, $0 < y_f(t) < 1$, $0 < y_o(t) < 1$, and $y_v(t) + y_f(t) + y_o(t) = 1.0$,

$$f\{y_v(t), y_f(t), y_o(t)\} = \frac{\Gamma(\alpha_{v,t} + \alpha_{f,t} + \alpha_{o,t})}{\Gamma(\alpha_{v,t})\Gamma(\alpha_{f,t})\Gamma(\alpha_{o,t})} \{y_v(t)\}^{\alpha_{v,t}-1} \{y_f(t)\}^{\alpha_{f,t}-1} \{y_o(t)\}^{\alpha_{o,t}-1}. \quad (3)$$

While (3) describes what is only a 2−dimensional distribution, for the purpose of examining how possible models represent the problem it is convenient to consider the implications

of those models for each of the three components, keeping in mind the restriction on their sum. For the model of expression (3) expected values are,

$$\begin{aligned} E\{Y_v(t)\} &= \frac{\alpha_{v,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \\ E\{Y_f(t)\} &= \frac{\alpha_{f,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \\ E\{Y_o(t)\} &= \frac{\alpha_{o,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}. \end{aligned} \tag{4}$$

In the notation developed in this question, the modeling strategy of Hijazi and Jernigan (2009) would be to take $(Y_v(t), Y_f(t), Y_o(t))$ to be independent over time t, and further model

$$\begin{aligned} \log(\alpha_{v,t}) &= \beta_{v,0}+\beta_{v,1}z_t, \\ \log(\alpha_{f,t}) &= \beta_{f,0}+\beta_{f,1}z_t, \\ \log(\alpha_{o,t}) &= \beta_{o,0}+\beta_{o,1}z_t, \end{aligned} \tag{5}$$

with unrestricted regression parameters $\beta_{v,0}, \beta_{f,0}, \beta_{o,0}, \beta_{v,1}, \beta_{f,1}$ and $\beta_{o,1}$.

ANSWER QUESTION 8 NOW. (SEE PAGE 16.)

At least for models based on multinomial distributions (which are sometimes considered discrete analogs of Dirichlet distributions), one common structure is to make use of "log odds" in building models. To apply this strategy to the problem of this question, let $\mu_{v,t}$, $\mu_{f,t}$ and $\mu_{o,t}$ denote the expected values given in equation (4) at time t, and take

$$\begin{aligned} \log\left(\frac{\mu_{v,t}}{\mu_{o,t}}\right) &= \beta_{v,0}+\beta_{v,1}z_t = \eta_{v,t} \\ \log\left(\frac{\mu_{f,t}}{\mu_{o,t}}\right) &= \beta_{f,0}+\beta_{f,1}z_t = \eta_{f,t}. \end{aligned} \tag{6}$$

Note that (6) implies that

$$\begin{aligned} \mu_{v,t} &= \frac{\exp\{\eta_{v,t}\}}{1+\exp\{\eta_{v,t}\}+\exp\{\eta_{f,t}\}}, \\ \mu_{f,t} &= \frac{\exp\{\eta_{f,t}\}}{1+\exp\{\eta_{v,t}\}+\exp\{\eta_{f,t}\}}, \\ \mu_{o,t} &= \frac{1}{1+\exp\{\eta_{v,t}\}+\exp\{\eta_{f,t}\}}. \end{aligned} \tag{7}$$

ANSWER QUESTION 9 NOW. (SEE PAGE 17.)

One potential structure for the regression model that would produce interpretable regression parameters would be to let

$$\begin{aligned} \mu_{v,t} &= \frac{\alpha_{v,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \\ \mu_{f,t} &= \frac{\alpha_{f,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \\ \mu_{o,t} &= \frac{\alpha_{o,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \end{aligned} \tag{8}$$

and then take

$$\begin{aligned} \log\left(\frac{\mu_{v,t}}{1-\mu_{v,t}}\right) &= \beta_{v,0}+\beta_{v,1}z_t = \eta_{v,t}, \\ \log\left(\frac{\mu_{f,t}}{1-\mu_{f,t}}\right) &= \beta_{f,0}+\beta_{f,1}z_t = \eta_{f,t}, \\ \log\left(\frac{\mu_{o,t}}{1-\mu_{o,t}}\right) &= \beta_{o,0}+\beta_{o,1}z_t = \eta_{o,t}. \end{aligned} \tag{9}$$

Another possible structure would be to let

$$\begin{aligned} \mu_{v,t} &= \frac{\alpha_{v,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \\ \mu_{f,t} &= \frac{\alpha_{f,t}}{\alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}}, \\ \text{and define } \phi &= \alpha_{v,t}+\alpha_{f,t}+\alpha_{o,t}, \end{aligned} \tag{10}$$

and then further model

$$\begin{aligned} \log\left(\frac{\mu_{v,t}}{1-\mu_{v,t}}\right) &= \beta_{v,0}+\beta_{v,1}z_t = \eta_{v,t}, \\ \log\left(\frac{\mu_{f,t}}{1-\mu_{f,t}}\right) &= \beta_{f,0}+\beta_{f,1}z_t = \eta_{f,t}. \end{aligned} \tag{11}$$

ANSWER QUESTION 10 NOW. (SEE PAGE 17.)

In an entirely exploratory mode, simple linear regression models were fit using ordinary least squares to the logits of observed vegetable and fruit acre proportions,

$$\begin{aligned} \log\left(\frac{y_v(t)}{1-y_v(t)}\right) &= \alpha_{v,0}+\alpha_{v,1}z_t+\sigma\epsilon_t, \\ \log\left(\frac{y_f(t)}{1-y_f(t)}\right) &= \alpha_{f,0}+\alpha_{f,1}z_t+\sigma\epsilon_t. \end{aligned} \tag{12}$$

Plots of fitted versions of these relations are presented in Figure 10. This exploratory analysis causes two concerns. First, the time regions in which residuals are positive and negative may have some similarlity betweeen vegetables and fruits (i.e., early 1960s, around 1990), although this pattern is not entirely clear (i.e., late 1990s). This is a concern because the Dirichlet distribution of expression (3) implies that the proportions of vegetables, fruits, and other crops should all have negative covariances. The correlation between residuals from the vegetable fit and the fruit fit is 0.261. A scatterplot of these residuals is presented in Figure 11, from which it can be seen that the weak positive correlation may be due to an absence of residual pairs that are negative for vegetables and positive for fruits. A second concern that the exploratory analysis of Figure 10 raises is that there appears to be evidence of the same autocorrelated error structure that was the focus of the analysis using raw vegetable acreage values (e.g., Figure 5).

ANSWER QUESTIONS 11 AND 12 NOW. (SEE PAGE 17.)

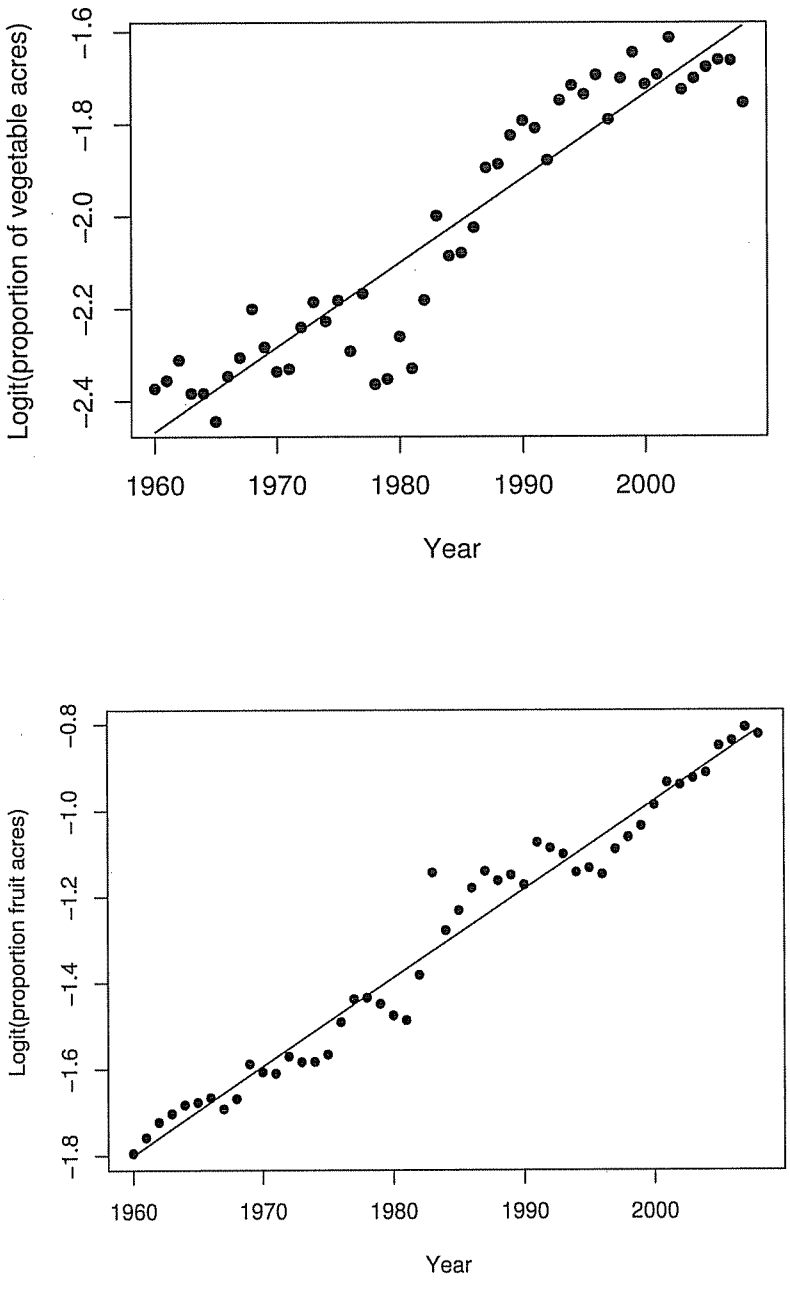


Figure 10: Logit proportions of vegetables and fruit over time.

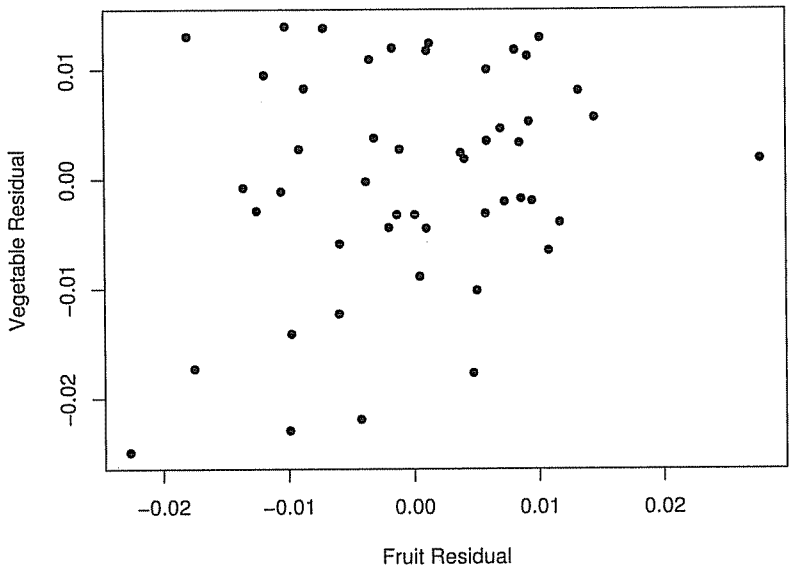


Figure 11: Scatterplot of residuals from the two regressions of Figure 10.

Questions

1. Assume that $E\{X(t)\} = 0$ and that second-order stationarity holds for the error process $\{X(t) : t = \ldots, -2, -1, 0, 1, 2, \ldots\}$. Derive the variance and covariance function for the response variables $Y(t)$.

   Hint: Second-order stationarity implies that $var\{X(t)\} = var\{X(t \pm 1)\}$ *for any* t *and* $cov\{X(t), X(t \pm k)\}$ *depends on* k *but not* t.

2. Suppose we would like to use maximum likelihood estimation for the parameters of the model specified in expressions (1) and (2).

   (a) Derive the log likelihood function for this model based on a sample of $\{Y(t) : t = 1, \ldots, n\}$.

   Hint: The model implies that, for any $t > 1$,

   $$f\{y(t)|y(1), \ldots, y(t-1)\} = f\{y(t)|y(t-1)\}.$$

   (b) You are given first derivatives of the log likelihood function as

$$\begin{aligned}
\frac{\partial}{\partial \beta_0}\ell(\gamma, \sigma^2, \beta_0, \beta_1) &= \beta_0 \left\{ -\frac{(1-\gamma^2)}{\sigma^2} - \frac{(n-1)(1-\gamma)^2}{\sigma^2} \right\} \\
&+ \beta_1 \left[-\frac{(1-\gamma^2)}{\sigma^2} z_1 - \frac{1-\gamma}{\sigma^2} \sum_{t=2}^{n} \{z_t - \gamma z_{t-1}\} \right] \\
&+ \frac{1-\gamma}{\sigma^2} \sum_{t=2}^{n} \{y(t) - \gamma y(t-1)\} + \frac{1-\gamma^2}{\sigma^2} y(1), \\
\frac{\partial}{\partial \beta_1}\ell(\gamma, \sigma^2, \beta_0, \beta_1) &= \beta_0 \left\{ -\frac{(1-\gamma^2)}{\sigma^2} z_1 - \frac{(1-\gamma)}{\sigma^2} \sum_{t=2}^{n} \{z_t - \gamma z_{t-1}\} \right\} \\
&+ \beta_1 \left[-\frac{(1-\gamma^2)}{\sigma^2} \{z_1\}^2 - \frac{1}{\sigma^2} \sum_{t=2}^{n} \{z_t - \gamma z_{t-1}\}^2 \right] \\
&+ \frac{1}{\sigma^2} \sum_{t=2}^{n} [\{y(t) - \gamma y(t-1)\}\{z_t - \gamma z_{t-1}\}] + \frac{1-\gamma^2}{\sigma^2} y(1) z_1, \\
\frac{\partial}{\partial \sigma^2}\ell(\gamma, \sigma^2, \beta_0, \beta_1) &= \frac{1-\gamma^2}{2\sigma^4} \{y(1) - \beta_0 - \beta_1 z_1\}^2 - \frac{n}{2\sigma^2} \\
&+ \frac{1}{2\sigma^4} \sum_{t=2}^{n} [\{y(t) - \gamma(t-1)\} - \beta_0(1-\gamma) - \beta_1 \{z_t - \gamma z_{t-1}\}]^2 \\
\frac{\partial}{\partial \gamma}\ell(\gamma, \sigma^2, \beta_0, \beta_1) &= \frac{\gamma}{\sigma^2} \{y(1) - \beta_0 - \beta_1 z_1\}^2 - \frac{\gamma}{1-\gamma^2} \\
&- \frac{\gamma}{\sigma^2} \sum_{t=2}^{n} [\{\beta_0 + \beta_1 z_{t-1}\}\{\beta_0 + \beta_1 z_{t-1}\}]
\end{aligned}$$

$$- \frac{1}{\sigma^2}\sum_{t=2}^{n}[\{y(t)-y(t-1)-\beta_0-\beta_1 z_t\}\{\beta_0+\beta_1 z_{t-1}\}].$$

With these derivatives and any other quantities you wish to derive, outline a procedure that could be used to obtain simultaneous maximum likelihood estimates of the parameters β_0, β_1, σ^2 and γ.

Hint: First write the derivatives in terms of additional summary quantities such as $T_1 = (1-\gamma^2) + (n-1)(1-\gamma)^2$, *for example, and then consider which parameter or parameters will pose the greatest challenge for estimation.*

3. There are a number of possible approaches for interval estimation with the model specified in (1) and (2), including intervals based on limiting distributions, and a number of parametric bootstrap intervals. Consider using bootstrap methods for interval estimation of the parameters of the model given by (1) and (2). Before we even compute interval estimates, why might we expect bootstrap percentile intervals to be preferable to basic bootstrap intervals?

4. Why do the autoregressive forecasts in Figure 6 and Figure 7 seem to differ from those produced using the model with independent errors (labeled OLS in Figure 7) for only a few future points in time?

5. Forecasts from any regression model (with a monotone systematic model component) for many time steps into the future are probably not appropriate in this problem. Explain this comment.

   Hint: The methodology used to produce forecasts from the autogressive model could be questioned, but that is NOT the issue we want to address in this question.

6. Now consider the procedure for producing forecasts described in the paragraph immediately following Table 2. Even if we produce a forecast for only one step ahead in time, the intervals that results from this procedure will be too narrow on average. Explain why this is the case. One sentence is sufficient to answer this question.

7. In this question consider forecasting only one time period into the future, that is, forecasting $Y(n+1)$, where $n = 2008$, the last observed time point. One approach for calculating a $(1-\alpha)100\%$ forecast interval would be to compute quantiles of the distribution of $Y(n+1)$ implied by the fitted model. If $F_{n+1}(y|y(n),\theta)$ denotes the conditional distribution function of $Y(n+1)$ implied by the model specified in (1)

and (2) given the observed value $Y(n) = y(n)$, and $\theta^T = (\beta_0, \beta_1, \sigma^2, \gamma)$ denotes the parameter vector of that model, we might then compute interval endpoints as

$$\begin{aligned} q_{\alpha/2} &= F_{n+1}^{-1}(\alpha/2|y(n), \hat{\theta}), \\ q_{1-\alpha/2} &= F_{n+1}^{-1}(1-\alpha/2|y(n), \hat{\theta}). \end{aligned} \tag{13}$$

It is known that the interval $(q_{\alpha/2}, q_{1-\alpha/2})$ will have actual coverage that differs from the nominal level of $1-\alpha$. A typical parametric bootstrap approach for estimating the actual coverage would be a follows.

(a) Simulate values $y_m^0(n+1)$ from the fitted model for $m = 1, \ldots, M$. These will become the values "to be predicted" and would be simulated from the distribution $F_{n+1}(y|y(n), \hat{\theta})$.

(b) Simulate bootstrap data sets $\boldsymbol{y}_m^* = \{y_m^*(t) : t = 1, \ldots, n\}$ also for $m = 1, \ldots, M$ from the fitted model.

(c) For each bootstrap data set, estimate θ as θ_m^*, and compute a forecast interval with endpoints $q_{(\alpha/2),m}^*$ and $q_{(1-\alpha/2),m}^*$ from (13) using θ_m^* and $y_m^*(n)$ in place of $\hat{\theta}$ and $y(n)$, respectively.

(d) Compute the actual coverage of this procedure as

$$1 - c(\alpha) = \frac{1}{M} \sum_{m=1}^{M} I\left(q_{(\alpha/2),m}^* \leq y_m^0(n+1) \leq q_{(1-\alpha/2),m}^*\right),$$

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

Briefly explain why the procedure outlined immediately above would NOT be appropriate in this problem. Suggest a possible (and simple) modification to overcome this difficulty.

8. Given a likelihood formed from the probability density functions (3), the regression model of (5) might fit the data adequately. But the model of expression (5) does not provide parameters that can be easily interpreted relative to the objective of analysis, which is to describe and quantify the patterns of change over time in proportional allocation of agricultural acres to crop categories. What is the difficulty with the regression model (5) relative to interpretation of parameters within the objective of the analysis?

Hint: This has nothing to do with the assumption of independence of $(Y_v(t), Y_f(t), Y_o(t))$

over time. Assume, until told otherwise, that such independence is a reasonable assumption.

9. Does use of the structures given in expressions (6) and (7) improve the situation with respect to interpretability of regression parameters $\beta_{v,1}$ and $\beta_{f,1}$?
   Hint: Focus only on $\beta_{v,1}$ *and note that* $\beta_{v,1} > 0$ *implies that* $\log(\mu_{v,t}/\mu_{o,t}) < \log(\mu_{v,t+1}/\mu_{o,t+1})$.

10. (a) Why would we be unable to conduct a meaningful analysis using the model of (8) and (9)?
    (b) The model of expressions (10) and (11) does not suffer the same deficiency as that of (8) and (9). But what might be a potential concern with this model?

11. If the model consisting of the Dirichlet densities (3), the parameterization (10) and the systematic model component (11) is estimated, resulting in $\hat{\beta}_{v,0}, \hat{\beta}_{f,0}, \hat{\beta}_{v,1}, \hat{\beta}_{f,1}$ and $\hat{\phi}$, outline a procedure that could be used to assess whether or not the model is adequate to describe the positive correlation of residuals in Figure 11. Do not be concerned with other possible model inadequacies, only the issue of positive covariance between the proportions of vegetable and fruit acres. Be concise but reasonably explicit about quantities that will be computed.

12. Assuming that we would like to retain the basic form of the model based on expressions (3), (10) and (11), suggest an extension of this model that might have the potential for dealing with the apparent temporal correlation exhibited in the exploratory plots of Figure 10. Do not worry about how you would conduct an analysis of your model, provide only the model structure itself.

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. Assuming $E\{X(t)\} = 0$ and $var\{X(t)\} = var\{X(t-1)\}$,

$$\begin{aligned} var\{X(t)\} &= var[E\{X(t)|X(t-1)\}] + E[var\{X(t)|X(t-1)\}] \\ &= var[\gamma X(t-1)] + E[\sigma^2] \\ &= \gamma^2 var[X(t)] + \sigma^2 \\ &= \frac{\sigma^2}{1-\gamma^2}. \end{aligned}$$

It is then necessary that $|\gamma| < 1$. To find the covariance,

$$\begin{aligned} cov\{X(t), X(t+1)\} &= E\{X(t)X(t+1)\} = E[X(t)\{\gamma X(t) + \epsilon(t+1)\}] \\ &= \gamma E\{X(t)X(t)\} = \gamma var\{X(t)\} = \gamma \frac{\sigma^2}{1-\gamma^2}. \end{aligned}$$

Similarly,

$$\begin{aligned} cov\{X(t), X(t+2)\} &= E\{X(t)X(t+2)\} = E[X(t)\{\gamma X(t+1) + \epsilon(t+2)\}] \\ &= \gamma E\{X(t)X(t+1)\} = \gamma cov\{X(t), X(t+1)\} = \gamma^2 \frac{\sigma^2}{1-\gamma^2}. \end{aligned}$$

Continuing in this fashion shows that

$$cov\{X(t), X(t+k)\} = \gamma^k \frac{\sigma^2}{1-\gamma^2}.$$

Since t is arbitrary and these covariances do not depend on t, then we also have that $cov\{X(t), X(t+k)\} = cov\{X(t), X(t-k)\}$ giving

$$cov\{X(t), X(t \pm k)\} = \gamma^k \frac{1}{1-\gamma^2}.$$

The response variables $Y(t)$ are linear functions of the $X(t)$. Hence,

$$\begin{aligned} E\{Y(t)\} &= \beta_0 + \beta_1 z_t \\ var\{Y(t)\} &= var\{X(t)\} = \frac{\sigma^2}{1-\gamma^2} \\ cov\{Y(t), Y(t \pm k)\} &= cov\{X(t), X(t \pm k)\} = \gamma^k \frac{1}{1-\gamma^2}. \end{aligned}$$

Question 2. (a) Using $f(\cdot)$ as a generic probability density function, the Markov property implies

$$f\{y(1),\ldots,y(n)\} = f\{y(1)\}f\{y(2)|y(1)\}f\{y(3)|y(2)\}\ldots f\{y(n)|y(n-1)\}.$$

Here, for the random variables $\{X(t) : t = 1,\ldots,n\}$,

$$f\{x(1)\} = \left(2\pi\frac{\sigma^2}{1-\gamma^2}\right)^{-1/2} \exp\left[-\frac{1-\gamma^2}{2\sigma^2}\{x(1)\}^2\right],$$

and, for $t = 2,\ldots,n$,

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

The log likelihood function for the error process is then,

$$\begin{aligned}\ell(\gamma,\sigma^2) &= (1/2)\log(1-\gamma^2) - (n/2)\log(2\pi\sigma^2) - \frac{1-\gamma^2}{2\sigma^2}\{x(1)\}^2 \\ &- \frac{1}{2\sigma^2}\sum_{t=2}^{n}\{x(t)-\gamma x(t-1)\}^2.\end{aligned}$$

With $Y(t)$ a linear function of $X(t)$, the log likelihood for the responses becomes

$$\begin{aligned}\ell(\gamma,\sigma^2,\beta_0,\beta_1) &= (1/2)\log(1-\gamma^2) - (n/2)\log(2\pi\sigma^2) - \frac{1-\gamma^2}{2\sigma^2}\{y(1)-\beta_0-\beta_1 z_1\}^2 \\ &- \frac{1}{2\sigma^2}\sum_{t=2}^{n}\{y(t)-\beta_0-\beta_1 z_t-\gamma y(t-1)+\gamma\beta_0+\gamma\beta_1 z_{t-1}\}^2 \\ &= (1/2)\log(1-\gamma^2) - (n/2)\log(2\pi\sigma^2) - \frac{1-\gamma^2}{2\sigma^2}\{y(1)-\beta_0-\beta_1 z_1\}^2 \\ &- \frac{1}{2\sigma^2}\sum_{t=2}^{n}[\{y(t)-\gamma y(t-1)\}-\beta_0(1-\gamma)-\beta_1\{z_t-\gamma z_{t-1}\}]^2.\end{aligned}$$

(b) First note that σ^2 cancels in solutions to $\partial\ell/\partial\beta_0 = 0$ and $\partial\ell/\partial\beta_1 = 0$. Then, following the hint, define

$$\begin{aligned}T_1 &= (1-\gamma^2) + (n-1)(1-\gamma)^2 \\ T_2 &= (1-\gamma^2)z_1 + (1-\gamma)\sum_{t=2}^{n}\{z_t-\gamma z_{t-1}\} \\ T_3 &= (1-\gamma^2)y(1) + (1-\gamma)\sum_{t=2}^{n}\{y(t)-\gamma y(t-1)\}\end{aligned}$$

$$
\begin{aligned}
T_4 &= (1-\gamma^2)\{z_1\}^2 + \sum_{t=2}^{n}\{z_t - \gamma z_{t-1}\}^2 \\
T_5 &= (1-\gamma^2)y(1)z_1 + \sum_{t=2}^{n}\{y(t) - \gamma y(t-1)\}\{z_t - \gamma z_{t-1}\} \\
S_1 &= (1-\gamma^2)\{y(1) - \beta_0 - \beta_1 z_1\}^2 \\
S_2 &= \sum_{t=2}^{n}[\{y(t) - \gamma y(t-1)\} - \beta_0(1-\gamma) - \beta_1\{z_t - \gamma z_{t-}\)\}]^2
\end{aligned}
$$

For a given value of γ, maximizing values in β_0, β_1, and σ^2 are available in closed form as

$$\hat{\beta}_0(\gamma) = \frac{1}{T_1}\left\{T_3 - T_1\hat{\beta}_1(\gamma)\right\}$$

$$\hat{\beta}_1(\gamma) = \frac{T_1T_5 - T_3T_2}{T_1T_4 - T_2^2}$$

$$\hat{\sigma}^2(\gamma) = \frac{1}{n}(S_1 + S_2)$$

where S_1 and S_2 are computed using values $\hat{\beta}_0(\gamma)$ and $\hat{\beta}_1(\gamma)$ in place of β_0 and β_1, respectively.

This suggests that the use of an unscaled profile likelihood procedure would lead to simultaneous maximum likelihood estimates as follows.

i. For a fixed value of γ define the log profile likelihood

$$\ell_p(\gamma) = \sup_{\beta_0,\beta_1,\sigma^2} \ell\{\gamma, \sigma^2, \beta_0, \beta_1\} = \ell\{\gamma, \hat{\sigma}^2(\gamma), \hat{\beta}_0(\gamma), \hat{\beta}_1(\gamma)\}.$$

ii. Maximize $\ell_p(\gamma)$ over γ using any one-dimenisonal search algorithm such as an equal interval search.

iii. The result is $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$ and $\hat{\gamma}$ that maximize $\ell\{\gamma, \sigma^2, \beta_0, \beta_1\}$. That is,

$$\sup_{\gamma}\ell_p(\gamma) = \sup_{\gamma}\sup_{\beta_0,\beta_1,\sigma^2}\ell\{\gamma, \sigma^2, \beta_0, \beta_1\} = \sup_{\gamma,\beta_0,\beta_1,\sigma^2}\ell\{\gamma, \sigma^2, \beta_0, \beta_1\}.$$

Question 3. Bootstrap percentile intervals have the property of automatically preserving the parameter space, a property not shared by basic bootstrap intervals. This could

potentially be an issue for interval estimates of γ, which has a bounded parameter space $|\gamma| < 1$. Whether or not this issue would be realized would depend on the observed data and how close the estimated value of γ is to a boundary of its parameter space.

Question 4. The autoregressive model is of order 1 with covariances that die off as a power of the autoregressive parameter $|\gamma| < 1$ (see question 1). The autoregressive forecasts are all conditioned on only $X(n)$ taken as equal to its surrogate value $y(n) - \hat{\beta}_0 - \hat{\beta}_1 z(n)$. Since this conditioning "wears off" rapidly after only a few time steps, simulated values for additional future times reflect marginal model behavior. The estimated marginal expected values for $Y(t)$ based on values from Table 2 are nearly identical to those based on values for the independence model of Table 1 (as they should be). Similarly, the marginal variance from the autoregressive model is $var\{Y(t)\} = \sigma^2/(1-\gamma^2)$ which, with estimates from Table 2 used as plug-in values, is 0.6975. The estimated variance from the independence model in Table 1 is 0.7100, again very similar values. Thus, the estimated marginal structure of the autoregressive model is essentially the same as that of the independence model.

Question 5. This comment concerns the fact that there are a finite number of acres available for agriculture in California. Figure 4 indicates that the total acres devoted to the three crop categories combined shows no definite trend from 1960 to 2008, although it has fluctuated considerably. This means that forecasts from any regression model with a monotone expectation function in time will eventually violate physical reality. Another way to say this is that the division of agricultural acres among the crop categories represents a "zero-sum game." Given that vegetable and fruit acres both have been increasing while acres in other crops has been decreasing, eventually forecasts from expected values that are linear over time will result in forecasted acres for vegetables and fruits that exceed the available land space and forecasts for other crops that become negative.

Question 6. The procedure outlined on page 5 ignores variability in parameter estimates.

Question 7. The difficulty with the parametric bootstrap procedure described in the question is that the future values "to be predicted" $y_m^0(n+1)$; $m = 1, \ldots, M$ are simulated from the actual fitted model, using the observed value $y(n)$, that is, from the distribution $F(y(n+1)|y(n), \hat{\theta})$. These values follow different conditional distributions than what would be used in the bootstrap replicates which is $F(y(n+1)|y_m^*(n), \theta_m^*)$. That is, $y_m^0(n+1)$ is conditioned on the last value from the actual data set, while forecasts from the bootstrap data sets would be conditioned on whatever the last value turns out to be in those data sets. The result that allows us to approximate probabilites with bootstrap procedures would not hold in this situation.

A potential solution to this difficulty would be to form intervals for bootstrap data sets using the distribution conditioned on the last value from the actual data but with the parameter value as estimated from the bootstrap data set, $F(y(n+1)|y(n), \theta_m^*)$.

Question 8. The difficulty with the model of expression (5) is that the regression coefficients $\beta_{v,1}$, $\beta_{f,1}$ and $\beta_{o,1}$ are not interpretable in terms of changes in the expected values of $Y_v(t)$, $Y_f(t)$ and $Y_o(t)$. It is possible, for example, that $\beta_{v,1} > 0$ and yet the expected proportion of acres devoted to vegetables will decrease over time. Similarly, it is possible that $\beta_{v,1} < 0$ and yet the proportion of vegetable acres increases over time. This is not merely a reversal of sign on coefficients, as increases are possible with $\beta_{v,1} > 0$ and decreases possible with $\beta_{v,1} < 0$ also. The problem lies with the fact that these regression coefficients are not related to expected values through a montone function.

Question 9. No, the structures in expressions (6) and (7) do not necessarily improve interpretation of $\beta_{v,1}$ and $\beta_{f,1}$ in this problem. For example, if $\beta_{v,1} > 0$ this implies that the log odds of vegetable acres relative to other crop acres increases, but this does not mean that the proportion of vegetable acres has necessarily increased. To see this, note that

$$\begin{aligned} & log\left(\frac{\mu_{v,t}}{\mu_{o,t}}\right) < \log\left(\frac{\mu_{v,t+1}}{\mu_{o,t+1}}\right) \\ \Rightarrow \quad & \log(\mu_{v,t}) - \log(\mu_{o,t}) < \log(\mu_{v,t+1}) - \log(\mu_{o,t+1}) \end{aligned}$$

$$\Rightarrow \log(\mu_{v,t}) - \log(\mu_{v,t+1}) < \log(\mu_{o,t}) - \log(\mu_{o,t+1})$$
$$\Rightarrow \frac{\mu_{v,t}}{\mu_{v,t+1}} < \frac{\mu_{o,t}}{\mu_{o,t+1}}$$

which can occur even if $\mu_{v,t} > \mu_{v,t+1}$. That is, even if the proportion of vegetable acres decreases (rather than increases), the odds and hence also log odds relative to the other crop category can still increase, if the decrease in this other crop category has been even greater than for vegetables.

Question 10. (a) The problem with a model having the structure of expressions (8) and (9) is that it is not identifiable. That is, given values for $\mu_{v,t}$, $\mu_{f,t}$ and $\mu_{o,t}$ there is not a unique solution for the parameters $\alpha_{v,t}$, $\alpha_{f,t}$ and $\alpha_{o,t}$. Any values of these fundamental parameters that are proportional will give the same expected values.

(b) A concern with the model of expressions (10) and (11) is how much the assumption that ϕ is constant for all times (all t) restricts the flexibility of the set of Dirichlet distributions being used in the model.

Question 11. Simulation based model assessment could be used to address this question. Here, we would implement the following steps:

(a) Using the estimated parameter values, compute, for $t = 1, \ldots, T$,

$$\begin{aligned}
\hat{\eta}_{v,t} &= \hat{\beta}_{v,0} + \hat{\beta}_{v,1} z_t, \\
\hat{\eta}_{f,t} &= \hat{\beta}_{f,0} + \hat{\beta}_{f,1} z_t, \\
\hat{\mu}_{v,t} &= \frac{\exp(\hat{\eta}_{v,t})}{1 + \exp(\hat{\eta}_{v,t})}, \\
\hat{\mu}_{f,t} &= \frac{\exp(\hat{\eta}_{f,t})}{1 + \exp(\hat{\eta}_{f,t})}.
\end{aligned}$$

(b) Then compute, for $t = 1, \ldots, T$,

$$\begin{aligned}
\hat{\alpha}_{v,t} &= \hat{\mu}_{v,t}\,\hat{\phi} \\
\hat{\alpha}_{f,t} &= \hat{\mu}_{f,t}\,\hat{\phi} \\
\hat{\alpha}_{o,t} &= \hat{\phi}\,(1 - \hat{\mu}_{v,t} - \hat{\mu}_{f,t}).
\end{aligned}$$

(c) Simulate, for $t = 1, \ldots, T$, and $m = 1, \ldots, M$, values $y_v^m(t)$, $y_f^m(t)$ and $y_o^m(t)$ from the Dirichlet densities of expression (3). If you know how to do this from composition of gamma variates so much the better, but that is not expected in this question.

(d) For $m = 1, \ldots, M$, fit regressions as in expression (12) using ordinary least squares, and compute the sample correlations among the two sets of residuals, r^m.

(e) With r_a denoting the actual correlation between the residuals from Figure 10, an assessment $p-$value is then given as

$$p = \frac{1}{M} \sum_{m=1}^{M} I(r^m \geq r_a).$$

An alternative assessment quantity that would also work here would be to compute the number of residual pairs that are negative for vegetables but positive for fruits.

Question 12. An extension of the model that could potentially deal with correlation over time would be to place a dynamic structure on the regression parameters $\beta_{v,1}$ and $\beta_{f,1}$. To accomplish this, let $\beta_{v,1}(t)$ and $\beta_{f,1}(t)$ denote these parameters at time t. Then model

$$\begin{aligned} \beta_{v,1}(t) &= \beta_{v,1}(t-1) + \lambda \epsilon_{v,t}, \\ \beta_{f,1}(t) &= \beta_{f,1}(t-1) + \psi \epsilon_{f,t}, \end{aligned} \tag{1}$$

where $\epsilon_{v,t} \overset{iid}{\sim} iidN(0,1)$ and $\epsilon_{f,t} \overset{iid}{\sim} iidN(0,1)$ and are independent.

Note: Although not expected in the answer, one potential for dealing with positive correlation between proportions of vegetable and fruit acres, should that prove to be a difficulty for the Dirichlet data model of expression (3), would be to take $\epsilon_{v,t}$ and $\epsilon_{f,t}$ to be correlated in this dynamic model structure.

Note: The data used in this prelim question may be found at

`www.nass.usda.gov/Statistics_by_State/California/Historical_Data/PCAcres.pdf`.