

Prelim Exam, Spring 1999

Stat 511

PROBLEM 1. Consider the linear model $E(\mathbf{Y}) = \mathbf{X}\beta$, $Var(\mathbf{Y}) = \sigma^2\mathbf{V}$, where \mathbf{Y} is an $n \times 1$ random vector, \mathbf{X} is a known $n \times p$ matrix of rank r , $r \leq p < n$, \mathbf{V} is a known $n \times n$ positive definite matrix, β is a $p \times 1$ vector of unknown parameters, and σ^2 is an unknown positive parameter.

- (i). Suppose that $\lambda'\beta$ is an estimable function. Give an expression for the vector ℓ (in terms of all or some of \mathbf{X} , β , σ^2 , \mathbf{V} and λ) such that $\ell'\mathbf{Y}$ is the best linear unbiased estimator (BLUE) of $\lambda'\beta$. (You do not have to justify your answer.)
- (ii). Suppose that the column space of the matrix $\mathbf{V}\mathbf{X}$ is contained in the column space of \mathbf{X} . Let $\lambda'\beta$ be an arbitrary estimable function. According to a result by Zyskind, the ordinary least squares (OLS) estimator of $\lambda'\beta$ should now be its BLUE. Show that if the stated condition for the column spaces holds, every solution to the normal equations is indeed also a solution to the Aitken equations.
- (iii). Consider the following two quadratic forms in \mathbf{Y} :

$$Q_1 = \mathbf{Y}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y},$$

$$Q_2 = \mathbf{Y}'(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1})\mathbf{Y}.$$

- (a). Show that $Q_2/(n - r)$ is an unbiased estimator of σ^2 .
- (b). Suppose that the column space of $\mathbf{V}\mathbf{X}$ is contained in the column space of \mathbf{X} . Can we now conclude that $Q_1/(n - r)$ is also an unbiased estimator of σ^2 ? Explain your answer.

PROBLEM 2. Let $\mathbf{Y} = (\mathbf{Y}_1', \dots, \mathbf{Y}_n')'$ be an $nt \times 1$ random vector, where, for $i = 1, \dots, n$, each \mathbf{Y}_i is a $t \times 1$ vector. Consider the model

$$\mathbf{Y}_i = \mathbf{X}\beta_i + \epsilon_i, \quad i = 1, \dots, n,$$

where \mathbf{X} is a known $t \times p$ matrix of rank p , the β_i 's ($p \times 1$) and ϵ_i 's ($t \times 1$) are random vectors with

$$\beta_i \sim N(\beta, \Sigma), \quad \epsilon_i \sim N(0, \sigma^2\mathbf{I}),$$

β is a $p \times 1$ vector of unknown parameters, Σ is an unknown $p \times p$ positive definite matrix, 0 is the $t \times 1$ vector of zeros, σ^2 is an unknown positive parameter, and \mathbf{I} is the $t \times t$ identity matrix. Assume that the $2n$ vectors $\beta_1, \dots, \beta_n, \epsilon_1, \dots, \epsilon_n$ are independently distributed.

Stat 511 - Page 2

- (i). Provide matrices Z and V (in terms of all or some of $n, t, p, X, \beta_i, \Sigma, \sigma^2$ and I) such that $Y \sim N(Z\beta, V)$.
- (ii). Let $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$. Show that β is estimable, and provide an expression for a matrix A (in terms of all or some of $n, t, p, X, \beta_i, \Sigma, \sigma^2$ and I) such that $A\bar{Y}$ is the OLS estimator of β .
- (iii). Show that for any $p \times 1$ vector λ the OLS estimator of $\lambda'\beta$ is also its BLUE, irrespective of what the positive definite matrix Σ and the positive parameter σ^2 are.
- (iv). Show that

$$\frac{1}{n(n-1)}(X'X)^{-1}X' \sum_{i=1}^n (Y_i - \bar{Y})(Y_i - \bar{Y})'X(X'X)^{-1}$$

is an unbiased estimator of the variance-covariance matrix of the OLS estimator of β .

- (v). For a known $p \times 1$ vector λ , give a constant d (in terms of some or all of n, t and p) and a matrix B (in terms of some or all of $n, t, p, X, \beta_i, \Sigma, \sigma^2$ and I) such that

$$\lambda'\hat{\beta} \pm t_{\alpha/2;d}\sqrt{\lambda'B\lambda},$$

where $\lambda'\hat{\beta}$ is the BLUE of $\lambda'\beta$ and $t_{\alpha/2;d}$ is the $100(1 - \alpha/2)$ -th percentile of the t -distribution with d degrees of freedom, is a $100(1 - \alpha)\%$ confidence interval for $\lambda'\beta$. Justify your answer.

- (vi). Construct a $100(1 - \alpha)\%$ confidence interval for σ^2 based on the statistic $\sum_{i=1}^n Y_i'(I - X(X'X)^{-1}X')Y_i$. Justify your answer.

Prelim Exam, Spring '99
 Stat 511 Solution

$$1 (i) \quad l = V^{-1}X(X'V^{-1}X)^{-1}\lambda$$

(ii) Since $\mathcal{C}(VX) \subset \mathcal{C}(X)$ and $\text{rank}(VX) = \text{rank}(X)$ (since V is nonsingular), it follows that $\mathcal{C}(VX) = \mathcal{C}(X)$. Hence $X = VX\Lambda$ for some Λ , or $V^{-1}X = X\Lambda$.

Let $\hat{\beta}$ be a solution to the NE, i.e. $X'X\hat{\beta} = X'Y$. Then also $\underbrace{A'X'X\hat{\beta}}_{X'V^{-1}} = \underbrace{A'X'Y}_{X'V^{-1}Y}$, or $X'V^{-1}X\hat{\beta} = X'V^{-1}Y$,

i.e. $\hat{\beta}$ is a solution to the Aitken equations.

(iii) (a) Recall (or derive) that $E(Y'A Y) = E(Y)'A E(Y) + \text{Tr}(A \text{Var}(Y))$. Here, for Q_2 ,

$$E(Y)'A E(Y) = \beta' X' V^{-1} X \beta - \beta' X' V^{-1} X (X' V^{-1} X)^{-1} X' V^{-1} X \beta \\ = 0$$

$$\text{Tr}(A \text{Var}(Y)) = \sigma^2 \text{Tr}(I - V^{-1}X(X'V^{-1}X)^{-1}X') \\ = \sigma^2 (n - \underbrace{\text{Tr}(X'V^{-1}X(X'V^{-1}X)^{-1})}_{\text{idempotent}})$$

$$= \sigma^2 (n - \text{rank}(X'V^{-1}X)) = \sigma^2 (n - \text{rank}(X)) \\ = \sigma^2 (n - r). \quad \text{The result follows.}$$

511-2

(iii) (b) No If $V = I$ then $Q_i/(n-r)$ is an unbiased estimator of σ^2 ; but, for example, if $V = 2I$ its expected value is $8\sigma^2$.

$$2(i) \quad Z = \begin{bmatrix} X \\ \vdots \\ X \end{bmatrix}_{n \times p}, \quad V = \begin{bmatrix} V_1 & \theta \\ \vdots & \ddots \\ \theta & V_p \end{bmatrix}, \text{ where}$$

$$V_i = X Z X' + \sigma^2 I$$

(ii) β is estimable if and only if all columns in Z are linearly independent. But

$$\text{rank}(Z) = \text{rank}(X) = p,$$

so that β is indeed estimable.

The NE are: $Z'Z\beta = Z'Y$, or

$$n X'X\beta = \sum_{i=1}^n X'Y_i$$

The unique solution (and OLS estimator of β) is

$$\hat{\beta} = \frac{1}{n} (X'X)^{-1} \sum_{i=1}^n X'Y_i = (X'X)^{-1} X' \bar{Y}$$

$$\text{Hence } A = (X'X)^{-1} X'$$

5II-3

(iii) The result follows (Zyshka) if $\mathcal{C}(VZ) \subset \mathcal{C}(Z)$.

$$VZ = \begin{bmatrix} V_x & \\ & V_y \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} X \sum X'X + \sigma^2 I \\ Y \sum X'X + \sigma^2 I \end{bmatrix}$$

$$= \begin{bmatrix} X \\ Y \end{bmatrix} (\sum X'X + \sigma^2 I) = Z (\sum X'X + \sigma^2 I),$$

so that indeed $\mathcal{C}(VZ) \subset \mathcal{C}(Z)$.

(iv) Let $W_i = (X'X)^{-1}X'(Y_i - \bar{Y})$. Then

$$\begin{aligned} E(W_i W_i') &= \text{Var}(W_i) + E(W_i) E(W_i)' = \\ &= \frac{n-1}{n} (X'X)^{-1} X' (\sum X'X + \sigma^2 I) X (X'X)^{-1} + 0 \\ &= \frac{n-1}{n} (\sum + \sigma^2 (X'X)^{-1}) \end{aligned}$$

Hence $E\left(\frac{1}{n(n-1)} \sum_{i=1}^n W_i W_i'\right)$

$$= \frac{1}{n} (\sum + \sigma^2 (X'X)^{-1}).$$

Since $\text{Var}(\hat{\beta}) = \text{Var}((X'X)^{-1}X' \bar{Y}) =$

$$(X'X)^{-1} X' \left(\frac{1}{n} (\sum X'X + \sigma^2 I) \right) X (X'X)^{-1} =$$

$\frac{1}{n} (\sum + \sigma^2 (X'X)^{-1})$, the result follows.

(v) From (iii) we know that $\hat{\lambda}'\hat{\beta}$ is the OLS estimator of $\lambda'\beta$.
Using (iv) and that γ is MVN, we see

$$\hat{\lambda}'\hat{\beta} \sim N(\lambda'\beta, \frac{1}{n} \lambda' (\sum + \sigma^2 (X'X)^{-1}) \lambda)$$

An unbiased estimator of this variance is (from (iv))

$$\frac{1}{n(n-1)} \sum_{i=1}^n X_i W_i W_i' \lambda ; \text{ this is a quadratic form in } \gamma, \text{ and}$$

we will denote it by Q . If we can show that

a $\hat{\beta}$ and Q are independently distributed, and

b $dQ/E(Q) \sim \chi_d^2$ for some d ,

$$\text{then } \frac{(\hat{\lambda}'\hat{\beta} - \lambda'\beta)/\sqrt{E(Q)}}{\sqrt{Q/E(Q)}} = \frac{\hat{\lambda}'\hat{\beta} - \lambda'\beta}{\sqrt{Q}} \sim t_d$$

The result follows then with the matrix B in the question equal to

$$B = \frac{1}{n(n-1)} \sum_{i=1}^n W_i W_i'$$

i.e. the matrix given in question (iv).

It remains to be shown that a and b hold, and in the process to determine the value of d .

For a, since γ is MVN, $\hat{\beta}$ is a function of γ

Q is a function of $Y_i - \bar{Y}$, $i=1, \dots, n$, it suffices to show that

$$\text{Cov}(\bar{Y}, Y_i - \bar{Y}) = 0, \quad i=1, \dots, n$$

And this is true since

$$\text{Cov}(\bar{Y}, Y_i) = \text{Var}(\bar{Y}) = \frac{1}{n}(X\sum X' + \sigma^2 I)$$

For b, one could write $Q = \gamma' M \gamma'$ for an appropriate symmetric matrix M , then determine a constant c so that $cM \text{Var}(\gamma)$ is idempotent, and then find $\text{rank}(M)$.

But perhaps easier, note that

$$\lambda'(X'X)^{-1}X'\gamma \sim N(\lambda'\beta, \lambda'(\Sigma + \sigma^2(X'X)^{-1})\lambda).$$

$$\text{Let } Z_i = \frac{\lambda'(X'X)^{-1}X'\gamma_i - \lambda'\beta}{\sqrt{\lambda'(\Sigma + \sigma^2(X'X)^{-1})\lambda}}$$

The Z_i 's are iid $N(0, 1)$, so that

$$\sum_{i=1}^n (Z_i - \bar{Z})^2 \sim \chi^2_{n-1}, \quad (\bar{Z} = \frac{1}{n} \sum_{i=1}^n Z_i)$$

$$\text{Since } Z_i - \bar{Z} = \frac{\lambda'(X'X)^{-1}X'(Y_i - \bar{Y})}{\sqrt{\lambda'(\Sigma + \sigma^2(X'X)^{-1})\lambda}} = \frac{\lambda'w_i}{\sqrt{n E(Q)}}$$

511-6

it follows that

$$\sum_{i=1}^n (z_i - \bar{z})^2 = (n-1) Q / E(Q)$$

Hence b holds with $d = n-1$.

(vi) Note that

$$(1) E(Y_i' (I - X(X'X)^{-1}X') Y_i) = \\ \text{Tr}((I - X(X'X)^{-1}X')(X \Sigma X' + \sigma^2 I)) \\ = \sigma^2 \text{Tr}(I - X(X'X)^{-1}X') = \sigma^2(t-p)$$

$$(2) Y_i' (I - X(X'X)^{-1}X') Y_i / \sigma^2 \text{ are iid } \chi_{t-p}^2$$

$$(3) \sum_{i=1}^n Y_i' (I - X(X'X)^{-1}X') Y_i / \sigma^2 \sim \chi_{n(t-p)}^2$$

$100\alpha/2$ th percentile

Hence

$$\Pr \left[\chi_{\alpha/2; n(t-p)}^2 \leq \frac{SS}{\sigma^2} \leq \chi_{1-\alpha/2; n(t-p)}^2 \right] = 1-\alpha$$

$100(1-\alpha)\%$ th percentile

where $SS = \sum_{i=1}^n Y_i' (I - X(X'X)^{-1}X') Y_i$. A $100(1-\alpha)\%$ confidence interval for σ^2 is thus given by

$$\left(\frac{SS}{\chi_{1-\alpha/2; n(t-p)}^2}, \frac{SS}{\chi_{\alpha/2; n(t-p)}^2} \right).$$

Methods I - March 1999 - Page 1 of 4

This question consists of two separate parts:

Part 1:

- a) Investigators want to compare a new treatment for leukemia to the standard. They randomly assign 6 subjects to each treatment group and compute their survival times (in weeks), which are given below. Notice that the trial ended after 15 weeks so those still alive at 15 weeks are denoted as 15+. Is the new therapy an improvement? Discuss various tests to use in this situation and which ones would be most appropriate; perform one of these 'appropriate tests' and give a p-value (exact if possible).

Standard Treatment: 4,5,6,7,9,10

New Treatment: 8,11,13,14,15+,15+

- b) If you knew that the survival times in each treatment group ($j = 1, 2$) were iid with density function $f(\cdot; \theta_j)$ and cdf $F(\cdot; \theta_j)$, write out the likelihood for estimating the parameters θ_j , $j = 1, 2$. In addition, based on the additional knowledge in this part, suggest and give the form for another test to determine if the new therapy is an improvement. Feel free to introduce additional notation.

Part 2:

Does pollution kill people? Data from an early study explored this issue in 60 Standard Metropolitan Statistical Areas (SMSA) in the U.S. from 1959-1961. The researchers wanted to examine the effects of two pollutants, oxides of nitrogen (NO_x) and sulfur dioxide (SO_2) after adjusting for several variables (given by X_1 , X_2 , and X_3 below). The variables are:

- Y = total age-adjusted mortality from all causes, in deaths per 100,000 population.
- X_1 = mean annual precipitation in inches
- X_2 = median number of school years completed, for persons of age 25 years or older
- X_3 = percentage of 1960 population that is non-white
- $X_4 = \log(NO_x)$
- $X_5 = \log(SO_2)$

Use the attached computer output to answer questions a), b), c).

- a) After adjusting for X_1 , X_2 , X_3 , as discussed above, is there a significant effect of both pollutants together? What about the effect of each pollutant given the other? Give test statistics, p-values, and your conclusions.

- b) Interpret the regression coefficients for X_4 and X_5 .

c) Does this model describe all the cities well? Notice that city 28 has a large studentized residual and city 37 has a large studentized residual and a large Cook's Distance. Explain how to compute the studentized residuals and how to assess whether an observation is really an outlier. Is city 37 an outlier? Give details. Also, explain how to compute Cook's Distance and describe what it actually measures. Can a city have a medium-sized residual and a large Cook's distance? Explain.

The next four parts are hypothetical situations given the above example:

d) Suppose you fit two regression models: one for the regression of Y on X_1, X_2, X_3, X_4 and the other for the regression of Y on X_1, X_2, X_3, X_5 . Is it possible to obtain estimates for the coefficients of X_4 and X_5 that are significantly different from zero when neither coefficient is significantly different from zero in the regression of Y on X_1, X_2, X_3, X_4, X_5 ? Explain.

e) Suppose that the researchers wanted to predict the mortality for a particular city with covariate vector \mathbf{X}_0 , where $\mathbf{X}_0 = (X_{01}, X_{02}, \dots, X_{05})'$. However, they are worried about whether predicting the mortality will require extrapolation beyond the space of \mathbf{X} 's used to fit the model. Give a numerical measure to assess this and explain why this measure is helpful.

f) Suppose that researchers hypothesized that the variable humidity, X_6 might be confounding the pollution effects on mortality. Give details of a graphical approach to determine whether to add X_6 into the model linearly, quadratically, Also, give an appropriate test statistic to determine whether X_6 should be added at all. If the pollution effects were still significant after adding X_6 into the model, could we make the statement that pollution caused the increased mortality that we observed? Explain.

g) Suppose that it turned out that the observations were not in fact independent, but dependent (i.e., there were $n = 60$ cities, but those closer together were more similar than those farther apart). The investigator becomes quite concerned about the estimates of the pollution effects in the model and their significance. Should he/she be concerned? Explain. In addition, give two approaches to alleviate his/her concern (hint: fitting a new model or modifying output from original model).

Methods I - March 1999 - Page 3 of 4

Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	157115.29331	31423.05866	23.846	0.0001
Error	54	71157.80097	1317.73706		
C Total	59	228273.09428			
Root MSE		36.30065	R-square	0.6883	
Dep Mean		940.35677	Adj R-sq	0.6594	
C.V.		3.86031			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > T
INTERCEP	1	940.654078	94.05423723	10.001	0.0001
X_1	1	1.946728	0.70069609	2.778	0.0075
X_2	1	-14.664053	6.93784607	-2.114	0.0392
X_3	1	3.028954	0.66851849	4.531	0.0001
X_4	1	6.715969	7.39895004	0.908	0.3681
X_5	1	11.358143	5.29548654	2.145	0.0365

Variable	DF	Type I SS	Type II SS	Standardized Estimate	Squared Corr Type I
INTERCEP	1	53056251	131805	0.00000000	.
X_1	1	59256	10171	0.31249136	0.25958252
X_2	1	20492	5886.911316	-0.19927978	0.08977185
X_3	1	51163	27051	0.43442221	0.22413008
X_4	1	20142	1085.690243	0.12788236	0.08823644
X_5	1	6062.219752	6062.219752	0.27346650	0.02655687

Methods I - March 1999 - Page 4 of 4

Obs	Residual	Std Err	Student	-2-1-0 1 2				Cook's	
				Residual	Residual			D	Rstudent
1	-12.8513	35.649	-0.360					0.001	-0.3576
2	82.7133	35.665	2.319		****			0.032	2.4213
3	25.5812	34.463	0.742		*			0.010	0.7391
4	-19.2352	34.696	-0.554		*			0.005	-0.5508
5	28.8494	34.385	0.839		*			0.013	0.8367
6	-52.3428	32.906	-1.591		***			0.091	-1.6142
.
28	-102.2	34.321	-2.978		*****			0.175	-3.2275
.
37	95.3830	26.828	3.555			*****		1.750	4.0246
.
38	19.8479	35.383	0.561		*			0.003	0.5574
39	12.5811	35.329	0.356					0.001	0.3532
40	20.7842	34.814	0.597		*			0.005	0.5934
41	-17.0559	34.399	-0.496					0.005	-0.4923
.
58	-3.8157	34.755	-0.110					0.000	-0.1088
59	-51.3315	33.636	-1.526		***			0.064	-1.5456
60	2.4415	35.943	0.068					0.000	0.0673

Methods I Solution - March 1999

Part 1:

a) We could use a Wilcoxon test or a randomization (permutation test). If we use the Wilcoxon test, the sum of the ranks on treatment 1 is 23. We cannot use the t-test here as the data is not normal (and small sample size) and there are censored observations. The exact p-value is .0087. Thus, the new therapy is an improvement.

b)

$$L(\theta_1, \theta_2 | \mathbf{x}) = \prod_{j=1}^2 \prod_{i=1}^{n_j} (f(x_{ji}; \theta_j))^{\delta_{ji}} (1 - F(x_{ji}; \theta_j))^{1-\delta_{ji}}$$

where $\delta_{ji} = 1$ if we observe the i th survival time in the j th group and 0 otherwise. An alternative test would be to test whether $\theta_1 = \theta_2$ using a likelihood ratio test.

Part 2:

a) yes, significance of both pollutants. compute $F = \frac{SS_{reg}(X_1, X_2 | X_1, X_2, X_3)/2}{MS_{error}(X_1, X_2)} = \frac{(20142+6062)/2}{1317.7} = 9.94$ which gives a p-value < .5. Effect of SO2 given everything else is significant, but the effect of NOx given everything else is not significant.

b) The regression coefficient for SO2 corresponds to the change in mortality for a unit change in log SO2 given everything else in the model held constant (same for NOX).

c) The studentized residual is computed as the $e_i / s(e_i)$ where $e_i = Y_i - \hat{Y}_i$ and $s(e_i) = \sqrt{(MSE(1 - h_{ii}))}$. We can use a Bonferroni correction to assess whether it is really an outlier. In this case, $n = 60$, so if $|e_i| > t_{n-5, 1-.05/(2*60)} = 3.54$. So city 37 is really an outlier here.

Cook's Distance can be computed as $D_i = \sum_{j=1}^n \frac{(\hat{Y}_j - \hat{Y}_{j(i)})^2}{pMSE}$. It measures the influence of the i th case on all n fitted values. It can also be expressed as a function of the residual and the leverage. So a city with a medium-sized residual can have either a large or small Cook's distance based on the leverage for that city.

d) Yes, if both pollutants are highly correlated. Illustrates problem of multi-collinearity

e) One can compute the leverage for \mathbf{X}_0 and see where this lies in terms of the leverage for the \mathbf{X} 's in the model. So compute $\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0$.

f) Could look at partial regression plots to examine how to enter humidity into the model. If the pollution effects were still significant, the association between pollution and mortality is still supported. A cause and effect statement cannot be made because this is an observational study and other factors such as temperature might be confounding the observed association.

g) The pollution effects will still be unbiased, but the variances will be incorrect. Two approaches to alleviate would be to fit a model with (spatially) dependent errors and do generalized least squares or to use a sandwich estimator for the variance will be robust to misspecification of the covariance structure.

- A. The following field experiment was performed to examine the effects of between rows and within row spacing of plants on the yield of a certain species of melon. Two levels of the between rows spacing factor were used. The rows were either 1.5 meters apart or 2.1 meters apart. Four levels of the within row spacing were used. Within a row, melon plants were spaced 0.6, 0.9, 1.2, or 1.5 meters apart.

The between rows and within row spacing of plants determines the plant density, the number of plants per hectare (ha). Yield was measured as the weight of the harvested melons per unit area, in this case metric tons per hectare ($t \cdot ha^{-1}$). Increasing plant density may increase yield per hectare if plants at the higher density have access to sufficient resources like water and nutrients. If plant density becomes too high, however, yield per hectare may show little improvement or even decline.

This study was replicated in five different fields. Each field was divided into eight plots and the eight combinations of between rows and within row spacings were randomly assigned to the plots. A separate randomization was done in each field. Four inch seedlings were transplanted to each plot according to the assigned between rows and within row spacings. All plots received the same amount of the same type of fertilizer and the same irrigation. Yields averaged across the five replicates are shown in the following table along with information on plant densities.

Between Rows Spacing (m)	Within Row Spacing (m)	Plant Density (plants/ha)	Yield (t/ha)
1.5	0.6	10764	19
	0.9	7176	18
	1.2	5382	14
	1.5	4305	13
2.1	0.6	7688	13
	0.9	5125	12
	1.2	3845	10
	1.5	3075	9

The researchers proposed the following model for the yields:

$$Y_{ijk} = \mu + R_i + B_j + W_k + (BW)_{ijk} + \varepsilon_{ijk},$$

where

B_j corresponds to the j -th level of the between rows spacing factor
 $j = 1, 2$

W_k corresponds to the k -th level of the within row spacing factor,
 $k = 1, 2, 3, 4$

R_i is a random field effect, with $R_i \sim NID(0, \sigma_R^2)$, $i = 1, 2, 3, 4, 5$

ε_{ijk} is a random error with $\varepsilon_{ijk} \sim NID(0, \sigma_\varepsilon^2)$

and R_i is independent of any ε_{ijk} .

The corresponding ANOVA table is shown below:

Source of Variation	Degrees of Freedom	Sums of Squares	Mean Squares
Replicates (fields)	4	86	21.50
Spacing Between Rows (B)	1	250	250.00
Within Row Spacing (W)	3	170	56.67
BxW Interaction	3	10	3.33
Residuals (error)	28	84	3.00
Corrected Total	39	600	

- (i) Write out the formula for the linear \times linear interaction contrast for the between rows and within row spacing factors.
 - (ii) Compute the sum of squares for the contrast in part (i).
- B. A suggestion was made to model the mean yield for the study described in part (A) as a straight line function of plant density, i.e.,

$$E(Y_{ijk}) = \beta_0 + \beta_1(X_{jk} - \bar{X}_.)$$

where X_{jk} is the plant density for the j -th between rows spacing and k -th within row spacing, and $\bar{X}_. = \frac{1}{8} \sum_{j=1}^2 \sum_{k=1}^4 X_{jk} = 5920$ plants/ha.

In answering the following questions, display relevant formulas. You are not required to evaluate these formulas to obtain numerical values for estimators or test statistics.

- (i) Give a formula (or formulas) to show how you would use the information given in part (A) to estimate β_0 and β_1 .
- (ii) Give formulas for standard errors of the estimators of β_0 and β_1 .
- (iii) Show how to test the null hypothesis that expected yield has a straight line relationship to plant density against the alternative that

$$E(Y_{ijk}) = \gamma_0 + g(X_{jk} - \bar{X}_{..})$$

where $g(X_{jk} - \bar{X}_{..})$ is an arbitrary function of $(X_{jk} - \bar{X}_{..})$.

- C. A second study was done by a different group of researchers. This study was performed in two locations. Five fields (or blocks) were available at each location. Each field was subdivided into 4 plots. The four within row spacings between plants (0.6, 0.9, 1.2, 1.5 meters) were randomly assigned to the four plots within each field with a separate randomization used for each field. The 1.5 meter spacing between rows was used in all plots in one location, and the 2.1 meter spacing between rows was used in all plots at the other location. The assignment of row spacings to locations was done by tossing a fair coin. All plots received the same amount of the same fertilizer and the same level of irrigation.
- (i) Outline the analysis of variance you would use for the data from this study. Specify sources of variation and degrees of freedom. (You do not have to show formulas for sums of squares or mean squares.) Use arrows to show how F-tests would be constructed.
 - (ii) Briefly summarize the advantages and disadvantages of this study, if any, relative to the study described in part (A).

- D. Since the size of the melons affects the price at which farmers can sell them, a third study was done to examine the effects of plant density on the distribution of melon sizes. Ten different fields were used in this study. In every field rows were spaced 2.1 meters apart. The within row spacing between plants was varied to vary the plant density. Within row spacings of 0.3, 0.6, 0.9, 1.2, and 1.5 meters were used and each level was randomly assigned to two of the ten fields. Samples of 250 melons were taken from the thousands of melons in each of the ten fields. In this study the size of a melon was determined as its weight. Consequently, for each field you have observed weights for a sample of 250 melons.

- (i) Briefly describe the steps you would take in developing a model for describing how the distribution of weights of melons changes with plant density. We want a model that would provide an estimate of the distribution of weights for the plant density corresponding to any within row spacing between 0.3 meters and 1.5 meters, not just the 5 levels at which data were collected.
- (ii) Give an example of a reasonable model and briefly indicate how you would estimate the parameters in your model.

A. (i) The contrast is

$$\begin{aligned}\gamma &= -(3\mu_{11} + \mu_{12} + \mu_{13} + 3\mu_{14}) + (-3\mu_{21} - \mu_{22} + \mu_{23} + 3\mu_{24}) \\ &= -(-3BW_{11} - BW_{12} + BW_{13} + 3BW_{14}) + (-3BW_{21} - BW_{22} + BW_{23} + 3BW_{24})\end{aligned}$$

and the least squares estimate is

$$c = -(-3\bar{Y}_{11} - \bar{Y}_{12} + \bar{Y}_{13} + 3\bar{Y}_{14}) + (-3\bar{Y}_{21} - \bar{Y}_{22} + \bar{Y}_{23} + 3\bar{Y}_{24}) = 8$$

$$(ii) \quad SS_c = \frac{5c^2}{40} = \frac{5(8)^2}{40} = 8$$

$$B. (i) \text{ Ordinary least squares estimators are } b_1 = \frac{\sum_{j=1}^2 \sum_{k=1}^4 \bar{Y}_{jk}(x_{jk} - \bar{x}_{..})}{\sum_{j=1}^2 \sum_{k=1}^4 (x_{jk} - \bar{x}_{..})^2} = .00120657$$

$$\text{and } b_0 = \bar{Y}_{..} = 13.5.$$

(ii) Substituting $\bar{Y}_{jk} = \mu + \bar{R}_j + B_j + W_k + (BW)_{jk} + \bar{\epsilon}_{jk}$ into the formula for b_1 we see that the average of the random block effects, \bar{R}_j , cancels out of the formula, and

$$\text{Var}(b_1) = \frac{\sigma_e^2}{5 \sum_{j=1}^2 \sum_{k=1}^4 (x_{jk} - \bar{x}_{..})^2} = \frac{\sigma_e^2}{5(44097040)} \text{ which is estimated as}$$

$$S_{b_1}^2 = \frac{3}{5(44097040)} \text{ and the standard error of the estimate is } S_{b_1} = .00011665$$

The estimate of the intercept is not free of random block effects, and

$$\text{Var}(b_0) = \text{Var}(\bar{Y}_{..}) = \frac{\sigma_e^2 + 8\sigma_R^2}{40} \text{ and the standard error is}$$

$$S_{b_0} = \sqrt{\frac{\text{MS}_{\text{replicates}}}{40}} = \sqrt{\frac{21.5}{40}} = .73314$$

(iii) Reject the straight line model proposed in part B if

$$F = \frac{(SS_{B,\text{part A}} + SS_{W,\text{part A}} + SS_{BxW,\text{part A}} - SS_{\text{model},\text{part B}})/(7-1)}{MS_{\text{residuals},\text{part A}}} > F_{(6,28), \alpha}$$

Here, $F = 6.05 > 4.02 = F_{(6,28), .01}$ and the straight line model is rejected.

C. (i)	<u>Source of variation</u>	<u>degrees of freedom</u>
	Between rows spacing	1
	error (a) (among locations)	0
	error (b) (among fields within locations)	8
	within row spacing	3
	within x between row spacing interaction	3
	error (c)	24
	corrected total	39

- (ii) Since each spacing between rows was only used at only one location in the second experiment, there is no way to distinguish between location to location variation and a possible between rows spacing effect. This is a definite disadvantage of the second experiment.
 - For both experiments, contrasts corresponding to within row spacing effects or the interaction of between and within row spacing are estimated "within blocks". This is an attractive feature of each experiment.

 - D. (i) One might start by making a separate histogram of the size data for each field to examine the shape of the distributions and select a family of distributions that provides an adequate approximation to the size distribution at each plant density. Kernel estimators of density functions could also be used in this endeavor. Alternatively, one might try to use either of these methods to graphically determine if the family of normal distributions would be adequate for some simple transformation of the size data for each plant density. PROC INSIGHT in SAS, for example, would provide a useful interactive tool for the latter endeavor.
- Probability plots and goodness-of-fit tests could be used to determine if a proposed family of distributions provides an adequate model for this range of plant densities.
- If a suitable parametric family of distributions can be identified, the next step would involve estimation of the parameters at each of the five plant densities. Then, one would need to construct a model to describe how the parameter values change with plant density.
- (ii) Suppose the distribution of $Y = \log(\text{size})$ was found to be well approximated by a normal distribution for each of the five plant densities considered in the study. Let x denote the plant density. The mean and variance of Y may be related to plant density through the models

$$E(Y|x) = \alpha + \beta x \quad \text{Var}(Y|x) = \exp(\gamma + \delta x)$$

Maximum likelihood estimation could be used to obtain estimates of $(\alpha, \beta, \gamma, \delta)$. This is an outline of only one of many possible reasonable answers.

Methods Question 3

Individuals stopped by police for erratic or unsafe driving behavior may be given a 'breathalyzer' test to determine whether their blood alcohol level exceeds the legal limit of 0.10 percent by volume. Realizing that a breathalyzer test may not be able to exactly determine the blood alcohol level of a human subject, the state legislature put into law that a person is considered to be driving under the influence of alcohol if the reading from a properly administered breathalyzer test exceeds

$$0.10 + E$$

where E denotes a margin of error. The legislature did not specify how the margin of error should be determined.

The current procedure used to set a margin of error is as follows:

1. A number (exactly how many is not known) of breathalyzer readings are taken from a laboratory set-up using apparatus in which a mixture of water and alcohol (with 0.1 percent alcohol by volume) is vaporized.
2. The absolute deviations of readings from a value of 0.10 are computed.
3. The margin of error is defined as the average of the set of absolute deviations.

You are contacted by some authorities who are unsure whether this is the 'best' approach by which to compute a margin of error. You are asked to perform three tasks: (1) provide a critical evaluation of the current procedure used to calculate E , (2) comment on the use of the laboratory procedure by which measurements for the computation of E are obtained, and (3) suggest a more appropriate or improved procedure.

1. Comment on the current procedure. In your answer consider the following.
 - (a) What can be said about using the mean of the absolute deviations as a margin of error? Assume that (i) the measurements are accurate (i.e., unbiased), (ii) the true mean absolute deviation, say δ is known, and (iii) the true standard deviation of measurements, say σ is known. With Y denoting a measured value, compare the use of the criterion $Y + \delta$ with the criterion $Y + \sigma$.
 - (b) Is the intended use of a margin of error in this application consistent with the 'usual' statistical notion of such a quantity? Explain.

Methods Question 3, p. 2

2. Comment on the current measurement procedure using a laboratory apparatus. In your answer consider the following.
 - (a) Does the current measurement procedure, based on a laboratory apparatus, provide an adequate means of establishing a margin of error. What are the potential strengths and weaknesses of the current measurement procedure?
 - (b) Is there any alternative to the use of a laboratory apparatus such as currently used to obtain ‘true’ values? What might be the benefits, as well as limitations of any alternative you suggest?
3. Starting from scratch (i.e., ignoring what is currently done), offer a definition of ‘margin of error’ for the intended use of a breathalyzer, and design a procedure to compute that margin of error. In preparing your answer, make simplifying assumptions as needed, and indicate the following.
 - (a) What basic information would you like to have available concerning the distribution of measurements or the distribution of differences between measured and true values?
 - (b) Let Y denote a measured value and consider the model

$$Y = T + \epsilon,$$

for some random variable ϵ with $E(\epsilon) = 0$. Here, T represents the true blood alcohol level of a subject and ϵ represents the deviation of the measured and true values. Assuming that the distribution of ϵ may be taken as normal with constant variance among a set of ϵ_i , what would be a simple procedure by which to set a margin of error?

- (c) If the distribution of measured values could not be assumed symmetric about the true value, how might you set a margin of error?
- (d) Assuming that the variance of measured values does not depend on the magnitude of the true value, and that this variance is known (possibly on the basis of an extensive sample), can you think of any way to set a ‘margin of error’ that does not in any way depend on the distributional form of the measurements?
Hint: think of inequalities that might be applied with known variance.

Methods Question 3
Answer

1. Comments on the Current Procedure.

- (a) There is no statistical framework discernible from the description of the current procedure given. The procedure deals with the estimation of what is generally called the 'mean deviation'. Using this estimate as a margin of error is generous to the measuring device. To see this, let Y be a random variable with $E(Y) = \mu$ and $\text{var}(Y) = \sigma^2$. Also, let $E(|Y - \mu|) = \delta$. Then,

$$\begin{aligned}\text{var}(|Y - \mu|) &= E\{(|Y - \mu| - \delta)^2\} \\ &= E(|Y - \mu|^2) - \delta^2 \\ &= \sigma^2 - \delta^2.\end{aligned}$$

Thus, $\delta < \sigma$ and, even if δ were known exactly (rather than requiring estimation), the best that the procedure could do would be to result in a conclusion that true blood alcohol level was greater than 0.10 if a measured value was greater than $0.10 + \sigma$.

- (b) The intended use of a margin of error in this problem is complicated by the fact that, in application, we will have a sample of size $n = 1$. This differs from the usual statistical formulation in which a margin of error is attached to an estimated proportion or mean.

2. Comments on the Current Measurement Procedure.

- (a) An adequate measurement process is generally considered to consist of i) an object to measure, ii) a well defined characteristic of the object, and iii) an appropriate measurement tool. In the actual application of breathalyzer tests, the object to be measured is clear – drivers suspected of being under the influence of alcohol. Also in application, the characteristic to be measured is fairly well defined – blood alcohol over 0.10 percent by volume. In the procedure currently used to define and calculate a 'margin of error', however, neither of these components of the measurement process are present. Instead, the goal of obtaining control over potential confounding influences (e.g., gender, weight, body condition) and the need for a 'true' value against which to compare measurements has resulted in design of a laboratory apparatus that might cause the assessment of measurement accuracy and precision to be quite different than

what is actually realized in application. We are given no information about whether the laboratory apparatus used is considered to adequately reflect the physiological processes under which blood alcohol levels can be measured on the basis of alcohol concentration in breath. One question about the current procedure is whether the laboratory setting adequately reflects the measurement process as it is to be used in actual application.

A related concern is that potentially confounding factors, such as those listed in the previous paragraph, cannot be controlled in application. The current procedure provides no means by which the importance or unimportance of such factors for measurement validity and reliability may be determined.

- (b) A possible alternative to the use of a laboratory apparatus would be to compare measured values to other measurements generally accepted as more accurate and precise than the breathalyzer measurements. Such an approach might, for example, compare measured breathalyzer values with corresponding results from analysis of blood samples taken from a collection of human subjects.

A difficulty in this suggestion is that control over the magnitude of the 'true' value is lost. The benefits would include the ability to determine whether results differ for groups of people with the grouping based on potentially confounding factors.

3. Alternative Procedures.

- (a) The basic characteristics of distributions are location, scale (or variability), and general shape. To set an adequate margin of error we would like to have information on all of these basic characteristics of the measured values or differences between measured and true value at a series of known true values. Important questions would include:
- Are the measured values unbiased for the true values?
 - Does the variance of measured values depend on the magnitude of the true value?
 - Is the distribution symmetric or not?
 - Does the distribution depend on other characteristics of the measurement setting other than the true value, such as temperature, physical characteristics of the individual, etc.?
- (b) The simplest procedure in this setting would be to mimic the development of an interval estimate of a normal mean, but using a fixed value for variance obtained from an extensive sample. That development would give a 'margin of error' as defined by the level of confidence in a traditional statistical interval. For example, using the common definition of a 95% level of confidence with the

term margin of error, an individual would be designated as over the legal limit of 0.10 if the measured value Y were such that $Y > 0.10 + 1.96z_{1-\alpha}\hat{\sigma}$, where $z \sim N(0, 1)$ and $\hat{\sigma}$ is the estimated variance obtained from a sample.

- (c) Here, our concern would be primarily with the distribution of measured values for a true value of 0.10, assuming that no dramatic changes in general shape or variance of the distribution occur across a reasonable range of true values centered at 0.10. An adequate margin of error in this case might be set as the $1 - \alpha$ quantile of an estimated distribution of measured values. An extensive sample of measurements would need to be obtained and the distribution estimated by either (1) the empirical distribution function, (2) a nonparametric density estimator (e.g., kernel), or (3) a fitted parametric model (with great attention given to goodness of fit).
- (d) If $\text{var}(Y)$ is known (or assumed known on the basis of an extensive sample) and $E(Y) = T$, a procedure could be based on the Chebyshev Inequality as follows:

The Chebyshev Inequality gives that,

$$\Pr(|Y - T| \geq t) \leq \frac{\text{var}(Y)}{t^2}.$$

For a particular measured value, y say, compute $t = y - 0.10$. Thus, if the true value was $T = E(Y) = 0.10$, the probability of a measurement as large or larger than y would be less than $\frac{\text{var}(y)}{t^2}$. A 'margin of error' could then be formulated as a lower bound on this quantity.