Ph.D. Prelim Exam, Spring 2000 **Linear Models**

Suppose that $\underline{y}$ is an $n \times 1$ observable random vector that follows the model

$$\underline{y} = X\underline{\beta} + \underline{\epsilon},$$

where X is an $n \times p$ known matrix of rank p^*, $\underline{\beta}$ is a $p \times 1$ vector of unknown parameters, and $\underline{\epsilon}$ is an unobservable random vector whose distribution is $N(\underline{0}, \sigma^2 H)$, where σ^2 is an unknown positive parameter, and H is a known positive definite matrix. Let $\underline{\tau} = \Lambda'\underline{\beta}$ represent a $q \times 1$ vector of linearly independent estimable functions, and let $\hat{\underline{\tau}}$ be the best linear unbiased estimator (BLUE) of $\underline{\tau}$. Define $\hat{\underline{\beta}}$ to be a solution to $X'H^{-1}X\hat{\underline{\beta}} = X'H^{-1}\underline{y}$, and let $\hat{\sigma}^2 = \underline{y}'H^{-1}(\underline{y} - X\hat{\underline{\beta}})/(n - p^*)$.

For some of the following questions, you may wish to use the fact that, for any positive definite matrix K, there exists a nonsingular matrix R such that $K = R'R$.

(a) State the generalization of the Gauss-Markov theorem to this model.

(b) Show that $\hat{\underline{\tau}} \sim N(\underline{\tau}, \sigma^2 \Lambda'(X'H^{-1}X)^- \Lambda)$.

(c) Show that $X(X'H^{-1}X)^- X'H^{-1}X = X$.

(d) Show that $\hat{\sigma}^2$ and $\hat{\underline{\tau}}$ are independently distributed.

(e) Let $A_1, ..., A_k$ be $n \times n$ symmetric matrices of rank $n_1, ..., n_k$, respectively. Suppose that $n_1 + \cdots + n_k = n$ and that $A_1 H + \cdots + A_k H = I$, where I is the $n \times n$ identity matrix. Show that $\underline{y}'A_1\underline{y}, ..., \underline{y}'A_k\underline{y}$ are independently distributed.

(f) Let A be an $n \times n$ symmetric matrix of rank r such that H is a generalized inverse of A. Find constants c_1, c_2 and c_3 and an $n \times 1$ vector of constants $\underline{a}$ so that $c_1 + \underline{a}'\underline{y} + c_2\underline{y}'A\underline{y}$ has a central chi-square distribution with degrees of freedom c_3. Note that these constants may depend on the unknown parameters $\underline{\beta}$ and σ^2.

(g) Use the results of parts (b), (c), and (d) and derive a size-γ test of the null hypothesis $H_0 : \underline{\tau} = \underline{b}$ versus the alternative hypothesis $H_a : \underline{\tau} \neq \underline{b}$.

(h) Show that, if

$$H = I + XA + B\left[I - X(X'X)^- X'\right], \qquad (1)$$

for some $p \times n$ matrix A and $n \times n$ matrix B, then $\underline{\lambda}'(X'X)^- X'\underline{y}$ is a BLUE of $\underline{\lambda}'\underline{\beta}$ for every $\underline{\lambda}$ for which $\underline{\lambda}'\underline{\beta}$ is estimable.

(i) If $\underline{\lambda}'(X'X)^- X'\underline{y}$ is a BLUE of $\underline{\lambda}'\underline{\beta}$ for every $\underline{\lambda}$ for which $\underline{\lambda}'\underline{\beta}$ is estimable, does condition (1) necessarily hold for some $p \times n$ matrix A and $n \times n$ matrix B? Explain your answer.

Ph.D. Prelim Exam — Solutions

Spring 2000 — Linear Models

(a) For the given model, if $\underline{\lambda}'\underline{\beta}$ is estimable and $\hat{\underline{\beta}}$ is any solution to $X'H^{-1}X\hat{\underline{\beta}} = X'H^{-1}\underline{y}$, then $\underline{\lambda}'\hat{\underline{\beta}}$ is a linear unbiased estimator of $\underline{\lambda}'\underline{\beta}$, and the variance of $\underline{\lambda}'\hat{\underline{\beta}}$ is uniformly less than that of any other linear unbiased estimator.

(b) Since $\underline{y} \sim N(X\underline{\beta}, \sigma^2 H)$, then $\hat{\underline{\tau}} = \Lambda'\hat{\underline{\beta}} = \Lambda'(X'H^{-1}X)^{-}X'H^{-1}\underline{y}$ is normally distributed. Furthermore, $E(\hat{\underline{\tau}}) = \underline{\tau}$ and $Var(\hat{\underline{\tau}}) = \sigma^2\Lambda'(X'H^{-1}X)^{-}\Lambda$. Thus, $\hat{\underline{\tau}} \sim N(\underline{\tau}, \sigma^2\Lambda'(X'H^{-1}X)^{-}\Lambda)$.

(c) Since H is positive definite, so is H^{-1}. Thus $H^{-1} = S'S$ for some nonsingular matrix S. Then

$$X(X'H^{-1}X)^{-}X'H^{-1}X = S^{-1}\underbrace{(SX)\cdot[(SX)'(SX)]^{-}(SX)'(SX)}_{\overset{\shortparallel}{SX}}$$

$$= S^{-1}(SX) = X.$$

[Note that $A(A'A)^{-}A'A = A$ for any matrix A.]

(d) It follows from the Aitken equations

$$X'H^{-1}X\hat{\underline{\beta}} = X'H^{-1}\underline{y}$$

that $\hat{\sigma}^2 = (\underline{y} - X\hat{\underline{\beta}})'H^{-1}(\underline{y} - X\hat{\underline{\beta}})/(n-p^*)$, which is a function of $\underline{y} - X\hat{\underline{\beta}}$. Thus it suffices to show that $\underline{y}-X\hat{\underline{\beta}}$ and $\hat{\underline{\tau}}$ are independent. Since the joint distribution of $\underline{y}-X\hat{\underline{\beta}}$ and $\hat{\underline{\tau}}$ (as linear functions of $\underline{y}$) is mutivariate normal, we only need to show that $\operatorname{cov}(\hat{\underline{\tau}}, \underline{y}-X\hat{\underline{\beta}})=0$.

We have $\operatorname{cov}(\hat{\underline{\tau}}, \underline{y}-X\hat{\underline{\beta}})$

$$
\begin{aligned}
&=\operatorname{cov}\left(\Lambda'(X'H^{-1}X)^{-}X'H^{-1}\underline{y},\ (I-X(X'H^{-1}X)^{-}X'H^{-1})\underline{y}\right)\\
&=\left[\Lambda'(X'H^{-1}X)^{-}X'H^{-1}\right]\cdot\sigma^2 H\cdot\left[I-X(X'H^{-1}X)^{-}X'H^{-1}\right]'\\
&=\sigma^2\Lambda'(X'H^{-1}X)^{-}\cdot\left[X-\underbrace{X(X'H^{-1}X)^{-}X'H^{-1}X}_{\overset{\shortparallel}{X}\ \text{from part (c)}}\right]'\\
&=0.
\end{aligned}
$$

(e). Since H is positive definite, $H=R'R$ for some nonsingular matrix R. Let $\underline{z}=(R^{-1})'\underline{y}$. Then $\underline{z}\sim N(\underline{\mu},\sigma^2 I)$, where $\underline{\mu}=(R^{-1})'X\underline{\beta}$. For $i=1,\cdots,n$,

$\underline{y}'A_i\underline{y}=\underline{z}'RA_iR'\underline{z}\equiv\underline{z}'B_i\underline{z}$, where $B_i=RA_iR'$.

Note that, $B_1,\cdots,B_k$ are symmetric matrices of rank $n_1,\cdots,n_k$, respectively, and $B_1+\cdots+B_k$ $=R(A_1+\cdots+A_k)R'=RH^{-1}R'=I$. Thus, by Cochran's Theorem, $\underline{z}'B_1\underline{z},\cdots,\underline{z}'B_k\underline{z}$ are independently distributed. Hence $\underline{y}'A_1\underline{y},\cdots,\underline{y}'A_k\underline{y}$ are independently distributed.

(f). Since H is a generalized inverse of A, then $AHA=A$ and $(AH)(AH)=AH$, implying that AH is an idempotent matrix of rank r. Note further that $\frac{1}{\sigma}(\underline{y}-X\underline{\beta}) \sim N(\underline{0}, H)$. Thus

$$\left[\frac{1}{\sigma}(\underline{y}-X\underline{\beta})\right]'A\cdot\left[\frac{1}{\sigma}(\underline{y}-X\underline{\beta})\right]\sim\chi^2(r).$$ That is,

$$\frac{1}{\sigma^2}\underline{\beta}'X'AX\underline{\beta} - \frac{2}{\sigma^2}\underline{\beta}'X'A\underline{y} + \frac{1}{\sigma^2}\underline{y}'A\underline{y}\sim\chi^2(r).$$

By taking $c_1 = \frac{1}{\sigma^2}\underline{\beta}'X'AX\underline{\beta}$, $c_2=\frac{1}{\sigma^2}$, $c_3=r$, and $\underline{a} = -\frac{2}{\sigma^2}AX\underline{\beta}$, we have that

$$c_1 + \underline{a}'\underline{y} + c_2\underline{y}'A\underline{y}\sim\chi^2(c_3).$$

(g). First we show that $\frac{(n-p^*)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-p^*)$.

Note that $\frac{(n-p^*)\hat{\sigma}^2}{\sigma^2} = \underline{y}'(\sigma^2H)^{-1}\cdot[I-X(X'H^{-1}X)^-X'H^{-1}]\underline{y}$ and that $\underline{y}\sim N(X\underline{\beta},\sigma^2H)$. Since

$$(\sigma^2H)^{-1}[I-X(X'H^{-1}X)^-X'H^{-1}](\sigma^2H) = I-H^{-1}X(X'H^{-1}X)^-X'$$

is an idempotent matrix (by part (c)) of rank $n-p^*$, and $(X\underline{\beta})'\cdot(\sigma^2H)^{-1}\cdot[I-X(X'H^{-1}X)^-X'H^{-1}]=0$, then $\frac{(n-p^*)\hat{\sigma}^2}{\sigma^2}\sim\chi^2(n-p^*)$.

It then follows from parts (b) and (d) that, under Ho,

$$\frac{(\hat{\underline{\tau}}-\underline{b})'\cdot[\Lambda'(X'H^{-1}X)^-\Lambda]^{-1}(\hat{\underline{\tau}}-\underline{b})}{q\hat{\sigma}^2}$$

$$\sim F_{q,n-p^*}.$$

Thus a size-γ F test is to reject H_0 if

$$\frac{(\hat{\underline{\tau}}-\underline{b})'[\Lambda'(X'H^{-1}X)^{-}\Lambda]^{-1}(\hat{\underline{\tau}}-\underline{b})}{q\hat{\sigma}^2} > F_{\gamma:q,n-p^*}.$$

(h) It suffices to verify that $HX=XQ$ for some matrix Q. Since $X(X'X)^{-}X'X=X$, then

$$HX=X+XAX=X(I+AX) \text{ for } Q=I+AX.$$

(i) If $\underline{\lambda}'(X'X)^{-}X'\underline{y}$ is a BLUE of $\underline{\lambda}'\underline{\beta}$ for every $\underline{\lambda}$ for which $\underline{\lambda}'\underline{\beta}$ is estimable, then $HP_X=P_XH$, where $P_X=X(X'X)^{-}X'$. Thus

$$\begin{aligned} H &= I+(H-I)P_X+(H-I)\cdot(I-P_X) \\ &= I+P_X(H-I)+(H-I)\cdot(I-P_X) \\ &= I+X\cdot[(X'X)^{-}X'(H-I)]+(H-I)\cdot[I-X(X'X)^{-}X'] \\ &= I+XA+B[I-X(X'X)^{-}X'] \end{aligned}$$

for $A=(X'X)^{-}X'(H-I)$ and $B=H-I$.

That is, condition (1) necessarily holds.

The chemical industry invests a great deal of money in research and development. To study the new product development process in the chemical industry a survey of firms is carried out. Fifty firms respond to the survey giving the number of new products developed by the company (PROD) in the last year along with the research and development budget (RND, measured in thousands of dollars), the total sales (SALES, measured in millions of dollars), and a measure of diversity of product line (DIVERS, measured on a scale from 0 to 1).

(i) An initial analysis suggests taking the logarithm of the response. Because some companies don't produce any new products we define LOGPROD = LOG(PROD + 1). A simple linear regression is carried out with the resulting equation

$$\begin{array}{rcccc} \text{LOGPROD} & = & .405 & + & .00099\ \text{RND} \\ & & (4.04) & & (3.08) \end{array}$$

The number in parentheses below each coefficient is its t-statistic.

(a) Is there evidence of a relationship between LOGPROD and RND expenditures? State your null hypothesis, alternative hypothesis, and conclusion.

(b) Give a 95% confidence interval for the coefficient of RND in this regression.

(c) A company plans to increase its RND budget by 100 thousand dollars. What would you tell them to expect in terms of new product development? Give a quantitative answer.

(ii) The previous analysis ignores potentially important variables, the size and diversity of the company (which are measured by SALES and DIVERS). A multiple regression analysis adding these two variables to RND is carried out. The attached output includes descriptive statistics for each variable, scatterplots showing the relationships of pairs of variables, and some regression output including a residual plot.

(a) Which variables are significant at the .05 level?

(b) Somehow the mean squared error was not included with the output. Compute its value from the output provided.

(c) Note that when SALES and DIVERS are added to the model the sign of the estimated coefficient of RND changes from positive (in the simple regression) to negative. Seeing these results, a cost-conscious vice-president decides to cut RND expenditures by 50% claiming this will actually improve product development. Explain why this reasoning is flawed.

(iii) Diagnostics

(a) Three companies appear to have high leverage values. Explain how leverage is computed and what it measures.

(b) The residual plot is generally fine except there is an odd pattern, the points appear almost to lie on a series of downward sloping parallel lines. Explain this pattern.

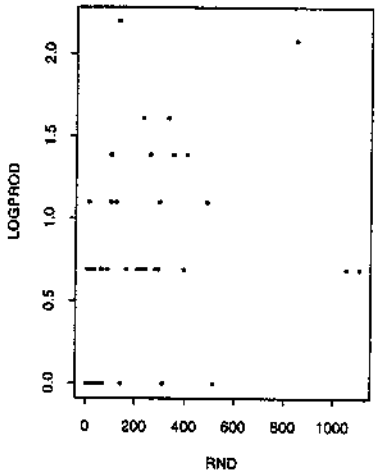
(iv) Missing data - Originally, the survey was mailed to 100 companies and only 50 responded. We analyzed the data from the 50 companies that responded in the previous parts of this question.

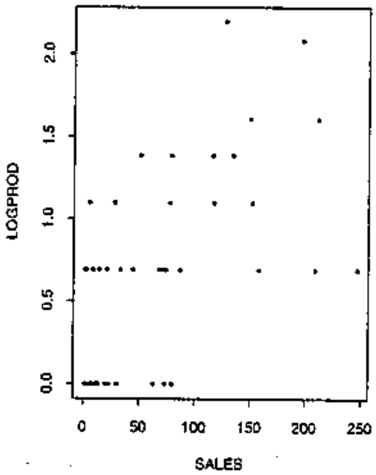
(a) Sales figures are available for all 100 companies because this is publicly available information. Based on the sales figures only, how might you tell if the missing companies were different from the responding companies? Explain in detail what statistical procedure/test you would use.

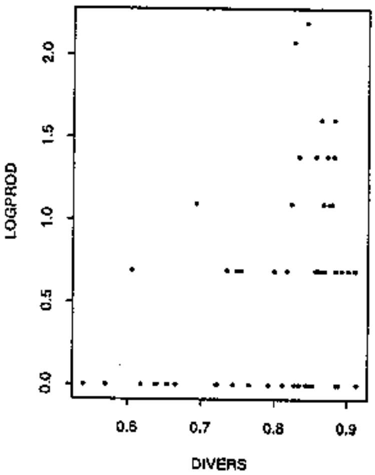
(b) The remaining 50 companies are contacted and 25 provide complete responses. How could you test for systematic differences in the product development process among the original respondents and the late respondents? Explain what statistical procedure/test you would use.

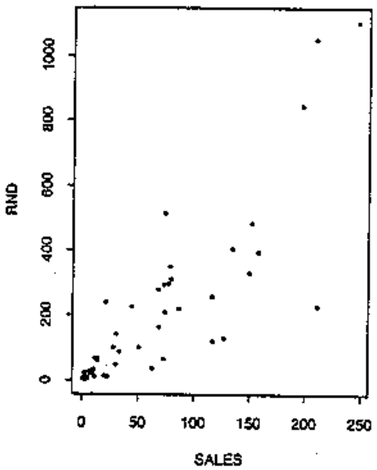
(v) Count data - One difficulty with these data are that the responses are integers ranging from 0 to 8 with 21 companies having 0 new products developed. Address this issue by describing an alternative method for analyzing these data to assess the relationship between SALES, RND, DIVERS and product development. Explain the method as well as possible, including a full specification of the model and how you would estimate the model parameters.

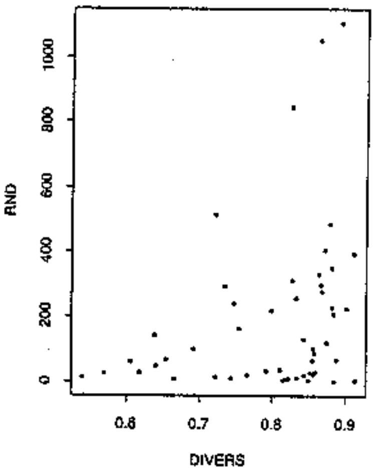
```
                          Simple Statistics                                    Correlation matrix
Variable      N    Mean   Std Dev  Min       Max             LOGPROD  RND   SALES DIVERS
LOGPROD      50    0.59      0.61    0      2.20               1.00   0.41   0.63   0.40
RND          50  190.28    250.71    2      1105               0.41   1.00   0.84   0.30
SALES        50   60.61     64.87  0.7     246.6               0.63   0.84   1.00   0.41
DIVERS       50    0.80      0.10  0.5       0.9               0.40   0.30   0.41   1.00

                   Parameter      Standard                                                Variance
  Variable   DF    Estimate          Error           Type I SS      Type II SS          Inflation
  INTERCEP    1   -0.448092    0.56948530           17.551788        0.128731         0.00000000
  RND         1   -0.001002    0.00048401            2.999452        0.890924         3.47004373
  SALES       1    0.008666    0.00196261            5.328695        4.054357         3.81981446
  DIVERS      1    0.883780    0.73944069            0.297028        0.297028         1.21987371

     Dep Var     Predict                    Student
Obs  LOGPROD       Value    Residual       Residual    CooksD    Rstudent     Leverage        Dffits
  1        0      0.3485     -0.3485         -0.795     0.013     -0.7918       0.0758       -0.2268
  2        0      0.3436     -0.3436         -0.768     0.006     -0.7649       0.0381       -0.1523
  3        0      0.3923     -0.3923         -0.875     0.007     -0.8730       0.0340       -0.1637
  4        0      0.3084     -0.3084         -0.697     0.007     -0.6927       0.0576       -0.1713
  5        0      0.3694     -0.3694         -0.841     0.014     -0.8384       0.0722       -0.2338
  6        0      0.1105     -0.1105         -0.266     0.004     -0.2635       0.1716       -0.1200
  7        0      0.1219     -0.1219         -0.280     0.002     -0.2775       0.0912       -0.0879
  8        0      0.3101     -0.3101         -0.725     0.018     -0.7214       0.1206       -0.2671
  9        0      0.3591     -0.3591         -0.809     0.009     -0.8056       0.0516       -0.1880
 10        0      0.3038     -0.3038         -0.678     0.004     -0.6738       0.0341       -0.1267
 11        0      0.2379     -0.2379         -0.543     0.006     -0.5391       0.0778       -0.1566
 12        0      0.1759     -0.1759         -0.399     0.003     -0.3956       0.0661       -0.1052
 13        0      0.6585     -0.6585         -1.464     0.015     -1.4829       0.0269       -0.2465
 14        0      0.0828     -0.0828         -0.195     0.001     -0.1932       0.1346       -0.0762
 15        0      0.7729     -0.7729         -1.737     0.038     -1.7777       0.0483       -0.4006
 16        0      0.3293     -0.3293         -0.753     0.012     -0.7496       0.0805       -0.2218
 17        0      0.1435     -0.1435         -0.325     0.002     -0.3217       0.0617       -0.0825
 18        0      0.2927     -0.2927         -0.659     0.006     -0.6545       0.0504       -0.1508
 19        0      0.3495     -0.3495         -0.786     0.008     -0.7826       0.0487       -0.1771
 20        0      0.2708     -0.2708         -0.604     0.003     -0.6000       0.0338       -0.1122
 21        0      0.3984     -0.3984         -0.917     0.021     -0.9156       0.0927       -0.2927
 22   0.6931      0.5056      0.1876          0.426     0.003      0.4220       0.0671        0.1132
 23   0.6931      0.1465      0.5467          1.264     0.045      1.2729       0.1009        0.4264
 24   0.6931      1.0706     -0.3775         -0.980     0.097     -0.9800       0.2869       -0.6217
 25   0.6931      0.7861     -0.0929         -0.207     0.000     -0.2045       0.0277       -0.0345
 26   0.6931      0.3508      0.3423          0.773     0.009      0.7694       0.0563        0.1880
 27   0.6931      0.8746     -0.1814         -0.409     0.002     -0.4055       0.0548       -0.0976
 28   0.6931      0.6261      0.0670          0.150     0.000      0.1482       0.0372        0.0291
 29   0.6931      0.7650     -0.0719         -0.161     0.000     -0.1588       0.0362       -0.0308
 30   0.6931      0.2748      0.4183          0.939     0.010      0.9377       0.0453        0.2042
 31   0.6931      1.3706     -0.6774         -1.769     0.327     -1.8122       0.2946       -1.1710
 32   0.6931      0.1485      0.5447          1.235     0.026      1.2423       0.0646        0.3264
 33   0.6931      0.6519      0.0412          0.092     0.000      0.0909       0.0314        0.0164
 34   0.6931      0.3016      0.3916          0.886     0.013      0.8841       0.0610        0.2253
 35   0.6931      0.5380      0.1552          0.347     0.001      0.3435       0.0373        0.0676
 36   0.6931      1.3259     -0.6327         -1.450     0.048     -1.4682       0.0843       -0.4455
 37   0.6931      0.5110      0.1821          0.407     0.002      0.4036       0.0384        0.0807
 38   1.0986      1.2174     -0.1188         -0.276     0.002     -0.2732       0.1082       -0.0952
 39   1.0986      0.6862      0.4124          0.920     0.007      0.9187       0.0341        0.1725
 40   1.0986      0.3020      0.7967          1.788     0.037      1.8328       0.0448        0.3968
 41   1.0986      1.1544     -0.0558         -0.126     0.000     -0.1250       0.0614       -0.0320
 42   1.0986      0.3113      0.7873          1.765     0.035      1.8084       0.0435        0.3858
 43   1.3863      1.0317      0.3546          0.797     0.008      0.7942       0.0488        0.1799
 44   1.3863      0.6588      0.7275          1.636     0.034      1.6673       0.0490        0.3783
 45   1.3863      0.6465      0.7398          1.650     0.024      1.6828       0.0335        0.3131
 46   1.3863      1.0769      0.3094          0.695     0.006      0.6915       0.0484        0.1560
 47   1.6094      1.9309     -0.3214         -0.882     0.110     -0.8796       0.3610       -0.6612
 48   1.6094      1.2812      0.3283          0.752     0.013      0.7484       0.0836        0.2261
 49   2.0794      1.1371      0.9423          2.262     0.253      2.3729       0.1652        1.0557
 50   2.1972      1.2624      0.9349          2.193     0.173      2.2922       0.1260        0.8704
```

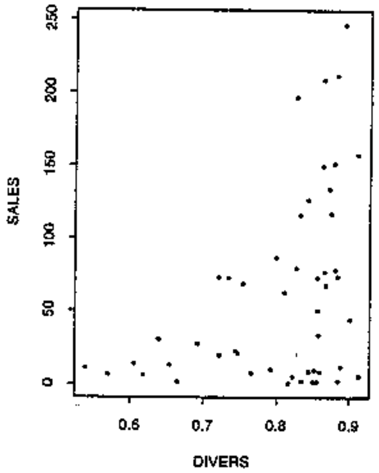

LOGPROD
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000
RND

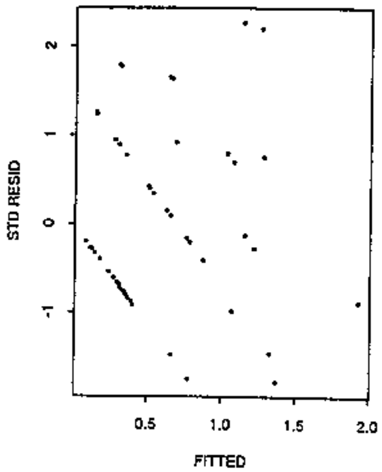

LOGPROD
0.0
0.5
1.0
1.5
2.0
0
50
100
150
200
250
SALES


LOGPROD
0.0
0.5
1.0
1.5
2.0
0.6
0.7
0.8
0.9
DIVERS


RND
0
200
400
600
800
1000
0
50
100
150
200
250
SALES


RND
0
200
400
600
800
1000
0.6
0.7
0.8
0.9
DIVERS


SALES
0
50
100
150
200
250
0.6
0.7
0.8
0.9
DIVERS


STD RESID
-1
0
1
2
0.5
1.0
1.5
2.0
FITTED

(i.a) There are 48 d.f.. The null hypothesis is that $\beta_{rnd} = 0$. For a two-sided alternative the p-value is .0034. Reject the null hypothesis at traditional (.05,.01) levels of significance.

(i.b) Standard error of $\hat{\beta}_{rnd}$ is $.00099/3.08 = .00032$. Then a 95% CI is $.00099 \pm 2.011 * .00032$ or $(.00035, .00163)$.

(i.c) According to the linear model we expect an increase in LOGPROD of approx .1. Exponentiating gives an increase in (PROD + 1) of about 10%. The addition of one is a nuisance here but basically we expect new products to increase by 10% or so.

(ii.a) The t-statistics are $-2.07, 4.42, 1.20$ from which it follows that the coefficients of RND and SALES are significantly different from zero at the .05 level.

(ii.b) There are many ways to get this. The least efficient is to calculate directly. Alternatives include the definition of studentized residual $r_i = e_i/\sqrt{MSE(1-h_{ii})}$ or the fact that partial (type II) SS divided by the MSE yields the square of the t-statistic for a coefficient. In either case $MSE \approx .208$.

(ii.c) SALES and RND are highly correlated. This correlation or collinearity can cause results like this. According to the multiple regression results an increase in RND expense of one thousand dollars keeping SALES and DIVERS fixed leads to a decrease in the expected amount of product development (one-tenth of one percent). This may reflect the fact that RND expenditures are not as effective in developing new products per dollar in large companies as they are in small companies. A key point is that it is difficult to justify the cost-cutting based on this observational study. A big company with no RND budget will likely not develop any new products at all despite what the regression model says.

(iii.a) If X is the matrix of predictor variables with a column of ones for the intercept, and 3 columns corresponding to RND, SALES, DIVERS, then h_{ii} is the ith diagonal of $X(X^TX)^{-1}X^T$. Informally h_{ii} is a measure of the distance from the ith row of X to the center of the point cloud, and hence a measure of potential inflence.

(iii.b) Notice that the response only takes on a few values. Residuals are computed as $Y - \hat{Y}$ so that all companies with the same Y value will have residuals on the same downward sloping line in a plot of residuals versus fitted values. It's not quite the case here because we have plotted studentized residuals.

(iv.a) One can do a two-sample t-test to compare the mean SALES in the responding and non-responding firms.

(iv.b) We now have 50 observations from the original respondents plus 25 observations from the group of follow-up respondents (sometimes known as the second wave). The question is whether the same regression surface fits both. You can introduce an indicator variable (and its interactions with the predictors) and carry out an F-test of the hypothesis that the corresponding coefficients are zero. Rejecting the null hypothesis suggests a different relationship in the two groups. This would be troubling because there are additional non-respondents out there.

(v) A natural idea is Poisson regression. If Y_i is the number of products developed, then our model would be $Y_i \sim \text{Poi}(\lambda_i)$ where the mean could be modeled as $\log(\lambda) = X\beta$. You can estimate the parameters by maximum likelihood. Another idea is logistic regression after dichotomizing the response.

Data were collected in the following manner in a study of the effects of nitrogen applications on the yield of corn. The experiment was replicated in six different fields, which we will refer to as blocks. Each field was divided into four strips. Each strip was 1500 feet long and wide enough to plant 36 rows of corn. The first treatment factor is the level of nitrogen applied to these strips in the spring before the corn is planted. We will call this factor the "spring application of nitrogen." It has four levels: 0, 50, 100, and 150 pounds per acre. These four levels were randomly assigned to the four strips within each field, with a separate randomization for each of the six fields.

The second factor is a second application of nitrogen applied in early summer when the corn plants are about two feet tall. We will refer to this factor as the "early summer application of nitrogen." It has three levels: 0, 50, and 100 pounds per acre. The levels of this factor are applied to sub-strips within each strip. Each strip is divided into three sub-strips with 12 rows of corn in each sub-strip. The three application levels of early summer nitrogen were randomly assigned to the three sub-strips within each strip, with a separate randomization within each of the 24 strips. All sub-strips were 1500 feet long.

At the end of the growing season, the corn was harvested and the total yield in bushels per acre was recorded for each of the 72 sub-strips. There were three sub-strips within each of the four strips in each of the six fields. The same variety of corn was planted in all of the sub-strips.

(a) Outline the ANOVA table you would use to analyze the effects of spring and early summer applications of nitrogen on corn yield. Present sources of variation and corresponding degrees of freedom.

(b) Using Y_{ijk} to represent the yield for the k-th level of the early summer application of nitrogen $(k = 1, 2, 3)$ and the j-th level of the spring application of nitrogen $(j = 1, 2\ 3, 4)$ in the i-th field, write out a formula for the linear model corresponding to your ANOVA in part (a). Describe the terms in your model, including any distributional assumptions or restrictions that you wish to impose. What does your model imply about correlations among corn yields from different sub-strips?

(c) To report a model in part (b) you had to make a decision about including the field effects as fixed effects or random effects. Describe the issues one should consider in deciding if block effects should be regarded as fixed effects or random effects. Describe the implications of this decision on the derivation of degrees of freedom, sums of squares and F-tests for the ANOVA table you reported in part (a). Describe the implications of this decision on the variances of the sample means

$$\overline{Y}_{\cdot jk} = \frac{1}{6}\sum_{i=1}^{6} Y_{ijk}\ , \quad j = 1, 2, 3, 4, \qquad k = 1, 2, 3$$

for the various combinations of spring and early summer applications of nitrogen.

(d) Comment on the potential advantages and disadvantages of the design of this experiment.

(e) Comment on the role of randomization in this experiment. Was it necessary or beneficial to randomly assign nitrogen application rates to strips and sub-strips? Explain, no credit will be given for simple "yes" or "no" answers.

(f) One objective of this study was to compare mean yields for the various combinations of spring and early summer levels of nitrogen application. Consider the sample means

$$\overline{Y}_{\cdot jk} = \frac{1}{6}\sum_{i=1}^{6} Y_{ijk}\ , \qquad j = 1, 2, 3, 4, \qquad k = 1, 2, 3$$

First describe how individual 95% confidence intervals can be constructed for differences of the form

$$E\left(\overline{Y}_{\cdot jk}\right) - E\left(\overline{Y}_{\cdot rs}\right).$$

Then address the issue of potentially making inferences for differences in 66 pairs of means.

(g) As the harvesting equipment moves along a 1500 foot sub-strip, it records corn yields in 15 foot intervals, producing 100 observations on yields within each sub-strip. Hence, 7200 observations on corn yields were recorded. There were two different soil types within each field which we will call types A and B. Using soil maps, the researchers could classify each 15 foot interval within a sub-strip into either the A or B soil category. Hence, yields inside 15 foot intervals could be matched to soil types. The researchers were interested in comparing the mean corn yields for a particular combination of spring and early summer applications of nitrogen (100 pounds per acre in spring followed by 50 pounds per acre in early summer) for the two types of soil. For this particular combination of nitrogen application rates, they computed the average $\overline{Y}_A$ and the sample variance s_A^2 of the $n_A = 387$ observations from the 15 foot intervals occurring in soil type A and the average $\overline{Y}_B$ and the sample variance s_B^2 of the $n_B = 213$ observations from 15 foot intervals occurring in soil type B. Then they computed the following statistic

$$t = \frac{\overline{Y}_A - \overline{Y}_B}{\sqrt{s^2\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}},$$

where

$$s^2 = \frac{(n_A - 1)\, s_A^2 + (n_B - 1)\, s_B^2}{n_A + n_B - 2}.$$

To test the null hypothesis that the soil type had no effect on corn yield, they compared t to the percentiles of a t-distribution with $n_A + n_B - 2$ degrees of freedom. Finally, they come to you for expert advice on whether or not this is a good way to proceed. How would you respond? Briefly state your reasoning.

(a) A simple split-plot analysis of variance is shown below. Would portioning the error(b) sum of squares into (block*early summer) and (block*spring*early summer) parts serve any useful purpose?

Source of variation	Degrees of freedom
Blocks (fields)	5
Levels of Spring nitrogen application	3
error (a) (strips within fields)	15
Levels of early Summer N application	2
Spring × early summer interaction	6
error (b)	40
corrected total	71

(b) A model that correspond to the ANOVA table shown in part (a) is

$$Y_{ijk} = \mu + \beta_i + \alpha_j + \eta_{ij} + \gamma_k + (\alpha\gamma)_{jk} + \varepsilon_{ijk}$$

where

α_j is a fixed effect associated with spring nitrogen application

γ_k is a fixed effect associated with early summer nitrogen application

$(\alpha\gamma)_{jk}$ is a fixed interaction effect

$\beta_i \sim \text{iid}(0, \sigma_\beta^2)$ are random block effects

$\eta_{ij} \sim \text{iid}(0, \sigma_\eta^2)$ are random effects associated with variation in corn yields among strips within fields that is not attributed to different levels of nitrogen application

$\varepsilon_{ijk} \sim \text{iid}(0, \sigma_\varepsilon^2)$ are random effects associated with variation in corn yields due to measurement errors and variation among sub-strips within strips that cannot be attributed to levels of nitrogen application.

Also, β_i, η_{ij}, and ε_{ijk} are assumed to be mutually independent. For this model, every observation has the same variance. Observations from different fields are uncorrelated. Observations from different strips in the same field have correlation $\sigma_\beta^2 / \left(\sigma_\beta^2 + \sigma_\eta^2 + \sigma_\varepsilon^2\right)$, and observations from different sub-strips in the same strip have correlation $(\sigma_\beta^2 + \sigma_\eta^2) / \left(\sigma_\beta^2 + \sigma_\eta^2 + \sigma_\varepsilon^2\right)$.

Including an additional set of random errors, $\delta_{ik} \sim \text{iid}(0, \sigma_\delta^2)$, in the model, i.e.,

$$Y_{ijk} = \mu + \beta_i + \alpha_j + \eta_{ij} + \gamma_k + (\alpha\gamma)_{jk} + \delta_{ik} + \varepsilon_{ijk},$$

allows for a slightly more complex correlation structure among the yields.

(c) If the fields were sampled from a larger population of fields that could have been used in this study, it would be reasonable to consider field effects as random block effects. Treating filed effects as fixed or random has no effect on the values of the sums of squares or the degrees of freedom in the ANOVA table. Variances of sample means can be affected, however, by the decision to use type of block effects included in the model. For the first model presented in part (b), for example,

$$\text{Var}(\bar{Y}_{\bullet jk}) = \frac{\sigma_\beta^2 + \sigma_\eta^2 + \sigma_\varepsilon^2}{6}$$

when blocks are random effects, but

$$\text{Var}(\bar{Y}_{\bullet jk}) = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2}{6}$$

when blocks are fixed effects.

(d) Contrasts among mean yields for various levels of early summer application of nitrogen and interaction contrasts involving spring and early summer applications are estimated more precisely than they would be if a randomized 4x3 factorial experiment had been performed within each field. Contrasts among mean yields for various levels of spring application of nitrogen are estimated with less precision.

(e) Comments on the role of randomization.

(f) In answering this question, note that $\text{Var}(\bar{Y}_{\bullet jk} - \bar{Y}_{\bullet rs}) = \dfrac{\sigma_\eta^2 + \sigma_\varepsilon^2}{3}$ when $j \neq r$, and

$\text{Var}(\bar{Y}_{\bullet jk} - \bar{Y}_{\bullet js}) = \dfrac{\sigma_\varepsilon^2}{3}$. (Hence, it is not reasonable to use an LSD approach as the researchers who collected the data wanted to do. Naive application of the GLM procedure in SAS can produce misleading results. The MIXED procedure in SAS can produce correct results.) Also address the issue of multiple comparisons.

(g) There are six hundred 15 foot segments of the sub-strips where a Spring application of nitrogen at 100 lb/acre was followed by an early summer application of 50 lb/acre. The suggested t-test is based on the assumption that the yield in any one of these segments is independent of the yield in any of the other 599 segments. This is not likely to be true. Yields in segments that are close together (and in the same sub-strip) may exhibit substantial positive correlation. How would this affect the proposed t-test? What can be done to account for these potential "spatial" correlations?

Ph.D Preliminary Exam
Spring 2000

Methods - 3

This problem explores three analyses of data on growth rates of guinea pigs. The investigators wanted to know whether vitamin E can reverse the effects of a growth inhibitor. Four weeks prior to the start of measurements, fifteen guinea pigs were given a dose of the growth inhibitor. One week prior to the first measurement, animals were randomly assigned to one of three groups (five guinea pigs per group). One group was given no Vitamin E, one group was given a low dose of Vitamin E, and the third was given a high dose of Vitamin E.

All animals were measured 1 week, 2 weeks, and 3 weeks after administration of the Vitamin E. In summary, there are 3 treatments, 15 guinea pigs and 45 observations. The data are plotted on the next page. Lines connect the 3 measurements on the same animal.

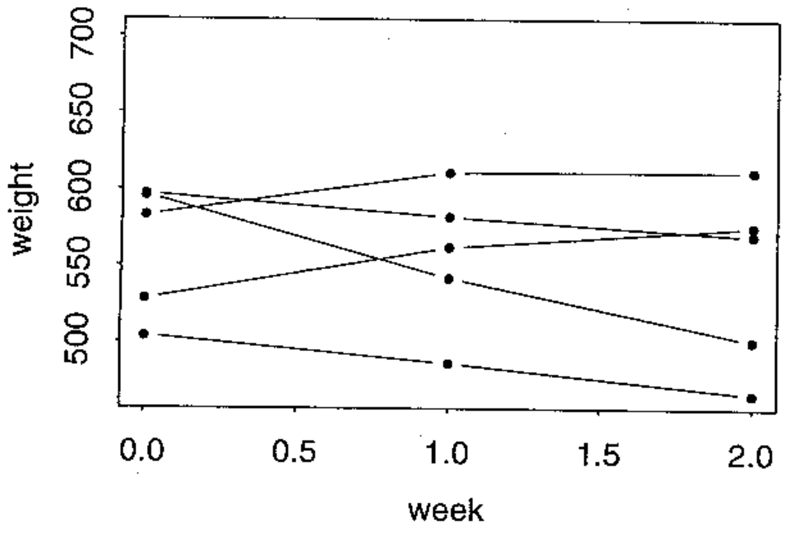
The researchers are interested in the growth rate of these guinea pigs, i.e. the slope of a linear regression of weight on week number. The objective of the study is to describe how vitamin E treatment affects the growth rate. Three important questions are:

1. Is the average growth rate the same in all three treatments?
2. What is the mean difference in average growth rate between the 'high vitamin E' and 'control' treatments?
3. How large is the between-individual variability in growth rate?

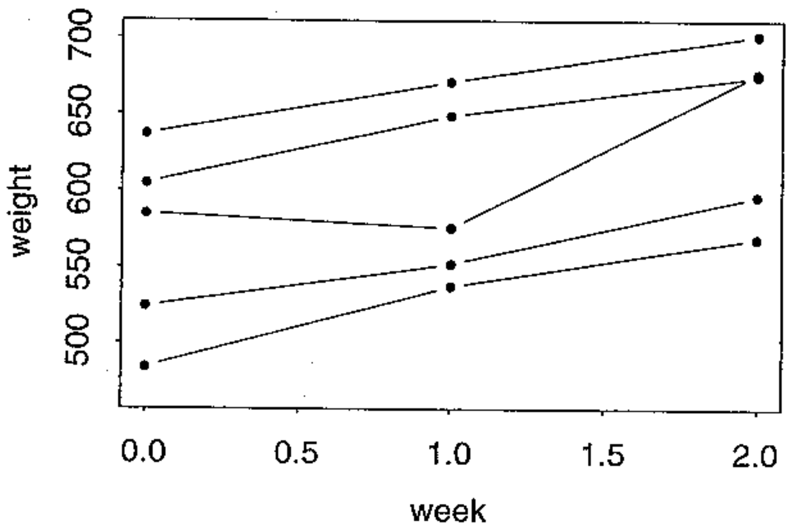
The three parts of this problem explore three approaches to answering these questions. In all three parts,

- the subscript i indicates treatment-specific quantities
- the subscript j indicates pig-specific quantities
- the subscript k indicates observation-specific quantities
- Y_{ijk}: weight of pig j in treatment i on week k
- X_{ijk}: week number (1, 2, or 3) for each observation.

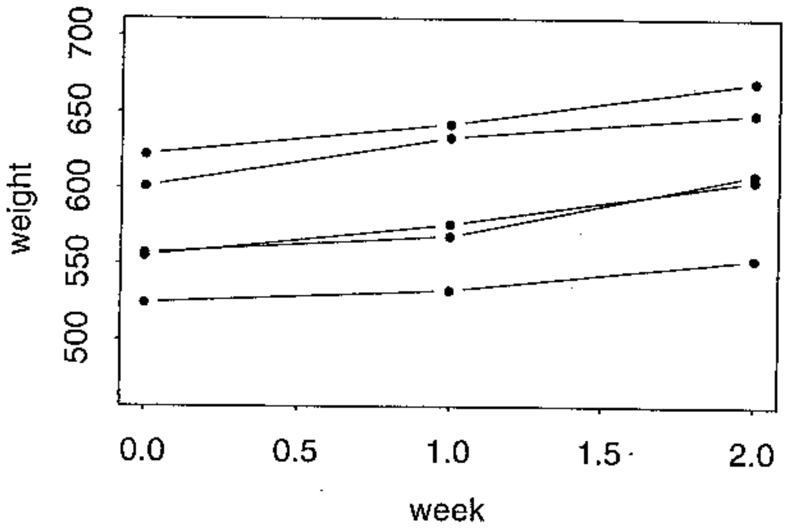
No Vitamin E



Low dose Vitamin E



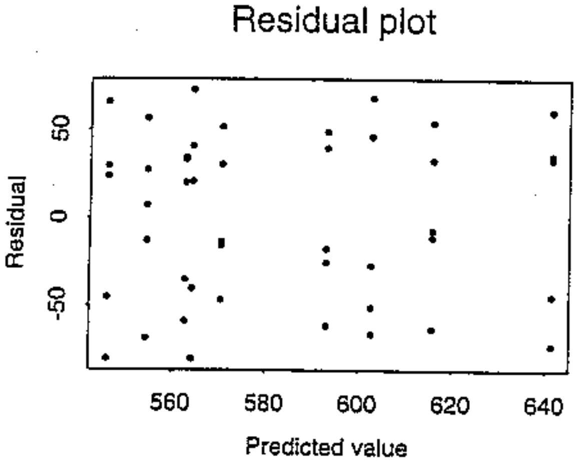
High dose Vitamin E

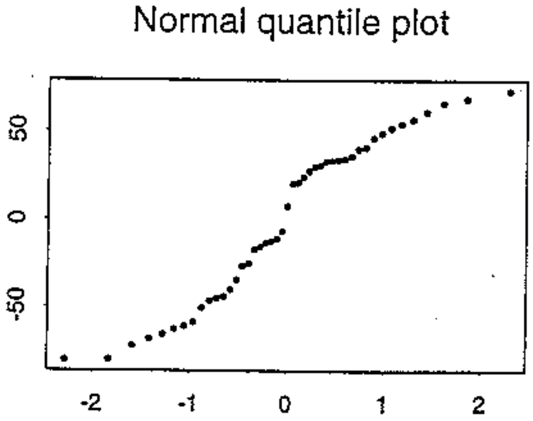


1. a) Four possible models that might be fit to the data are shown below with the residual SS for each. Use this information to construct a test of H$_0$: $\beta_1 = \beta_2 = \beta_3$ against the alternative that the slopes are not all equal. You should compute an appropriate test statistic and indicate how you would compute the p-value. You do not need to actually find the p-value.

Number	Model	Residual SS
1	$\text{E } Y_{ijk} = \mu$	139,076.6
2	$\text{E } Y_{ijk} = \alpha + \beta X_{ijk}$	123,022.0
3	$\text{E } Y_{ijk} = \alpha_i + \beta X_{ijk}$	106,655.3
4	$\text{E } Y_{ijk} = \alpha_i + \beta_i X_{ijk}$	101,012.1

b) Consider model 4, $\text{E } Y_{ijk} = \alpha_i + \beta_i X_{ijk}$. List the assumptions made by your test of $\beta_1 = \beta_2 = \beta_3$. Plotted below are a residual vs. predicted value plot and a normal quantile plot of the residuals. Use these diagnostics and the plot of the raw data to evaluate whether each of the assumptions is appropriate.





2. Another possible model is

$$Y_{ijk} = \alpha_{ij} + \beta_{ij} X_{ijk} + \varepsilon_{ijk}, \ \varepsilon_{ijk} \sim \text{ iid N}(0, \sigma^2)$$

i.e. a separate regression line is fit to each animal. The parameters are estimated by least squares. The residual Mean Square for this model is 515.27, with 15 d.f. Estimates of the slope, $\hat{\beta}_{ij}$, for each pig in the no Vitamin E group and the high Vitamin E group are:

Treatment	Pig	$\hat{\beta}_{ij}$	s.e. $\hat{\beta}_{ij}$
None	1	-19.0	9.406
None	2	-47.5	9.406
None	3	-13.5	9.406
None	4	14.5	9.406
None	5	24.0	9.406
High	11	24.0	9.406
High	12	26.0	9.406
High	13	25.0	9.406
High	14	24.0	9.406
High	15	14.5	9.406

Define $\overline{\beta}_i$ as the mean slope of all individuals receiving the i'th treatment.

a) Estimate $\overline{\beta}_{\text{High}} - \overline{\beta}_{\text{None}}$.

b) Test the hypothesis that $\overline{\beta}_{\text{High}} = \overline{\beta}_{\text{None}}$. Calculate a test statistic and indicate how you would compute the p-value, but you do not need to report a p-value.

c) Suppose you only assumed that the ε_{ijk} were independent with E $\varepsilon_{ijk} = 0$ and Var $\varepsilon_{ijk} = \sigma^2$. Could you still justify your estimator in part a) and test procedure in part b)? Why or why not?

3. A third possible model is the independent random coefficients model

$$\begin{aligned}
Y_{ijk} &= \alpha_{ij} + \beta_{ij} X_{ijk} + \varepsilon_{ijk} \\
\varepsilon_{ijk} &\sim \text{iid N}(0, \sigma^2) \\
\begin{bmatrix} \alpha_{ij} \\ \beta_{ij} \end{bmatrix} &\sim \text{independent N}\left(\begin{bmatrix} \overline{\alpha}_i \\ \overline{\beta}_i \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix} \right) \\
\sigma^2 &\geq 0 \\
\sigma_a^2 &\geq 0 \\
\sigma_b^2 &\geq 0 \\
\text{Cov}\,(\alpha_{ij}, \varepsilon_{ijk}) &= 0 \\
\text{Cov}\,(\beta_{ij}, \varepsilon_{ijk}) &= 0
\end{aligned}$$

a) Assume that the three variance components (σ^2, σ_a^2, and σ_b^2) are known. This model can be rewritten as the model

$$Y_{ijk} = \overline{\alpha}_i + \overline{\beta}_i X_{ijk} + \gamma_{ijk},$$

which can be written in vector form as

$$Y = X\beta + \gamma$$

where the elements of γ are no longer independent. Define Σ as the variance-covariance matrix of γ. Express Σ in terms of the variance components, σ^2, σ^2_a, and σ^2_b.

b) The elements of Σ are known when the three variance components, σ^2, σ^2_a, and σ^2_b, are known. In this case, what is a reasonable estimator for the vector of fixed effect parameters, β?

c) Consider estimating σ^2_b, the between-pig variance in slopes, using maximum likelihood. For these data (3 treatment groups), this model has 9 parameters: 3 mean intercepts, 3 mean slopes, and 3 variance components. Each variance component must be non-negative. One common approach to finding mle's is to set the score equations (the partial derivatives of the log likelihood with respect to each parameter) equal to zero, then solve that system of equations. In this model, does the solution to the score equations define the maximum likelihood estimate of the parameters? Why or why not?

d) Define $\hat{\sigma}^2_b$ as the mle of σ^2_b, the between-pig variance in slopes. Consider the likelihood ratio test of H$_0$: $\sigma^2_b = \sigma^2_0$ vs. the alternative $\sigma^2_b \neq \sigma^2_0$. For many choices of σ^2_0, the appropriate test statistic has an asymptotic Chi-square distribution with 1 d.f. The remaining parts of this question will evaluate whether this is the appropriate distribution for the likelihood ratio test of $\sigma^2_b = 0$.

For simplicity, reduce the problem to a 1 parameter problem by conditioning on specific values of the other eight parameters. Define:

$l(x)$ as the log-likelihood function evaluated at $\sigma^2_b = x$.
$D(x)$ as $-2(l(\sigma^2_0) - l(x))$.
σ^{2*}_b as the solution to the score equation, $\partial l / \partial \sigma^2_b = 0$.
τ^2 as equal to $-1/ \left.\frac{\partial^2 l(x)}{\partial x^2}\right|_{x=\sigma^{2*}_b}$.

First, we will imagine that there is no constraint on the parameter space for σ^2_b and connect the asymptotic distribution of $D(x)$ to the asymptotic distribution of an unconstrained estimate of σ^2_b.

Use a Taylor series expansion of $l(\sigma^2_0)$ around σ^{2*}_b to relate $D(\sigma^{2*}_b)$ to σ^{2*}_b and σ^2_0.

e) Relate the mle, $\hat{\sigma}^2_b$, to the unconstrained estimate, σ^{2*}_b, then use the approximation in d) to determine the asymptotic distribution of $D(\hat{\sigma}^2_b)$ under the null hypothesis, H$_0$: $\sigma^2_b = 0$. Identify the appropriate critical value for an asymptotic α=0.05 test of $\sigma^2_b = 0$ against the alternative: $\sigma^2_b > 0$.

Ph.D Preliminary Exam
Spring 2000

Methods - 3, with Answers

This problem explores three analyses of data on growth rates of guinea pigs. The investigators wanted to know whether vitamin E can reverse the effects of a growth inhibitor. Four weeks prior to the start of measurements, fifteen guinea pigs were given a dose of the growth inhibitor. One week prior to the first measurement, animals were randomly assigned to one of three groups (five guinea pigs per group). One group was given no Vitamin E, one group was given a low dose of Vitamin E, and the third was given a high dose of Vitamin E.

All animals were measured 1 week, 2 weeks, and 3 weeks after administration of the Vitamin E. In summary, there are 3 treatments, 15 guinea pigs and 45 observations. The data are plotted on the next page. Lines connect the 3 measurements on the same animal.

The researchers are interested in the growth rate of these guinea pigs, i.e. the slope of a linear regression of weight on week number. The objective of the study is to describe how vitamin E treatment affects the growth rate. Three important questions are:

1. Is the average growth rate the same in all three treatments?
2. What is the mean difference in average growth rate between the 'high vitamin E' and 'control' treatments?
3. How large is the between-individual variability in growth rate?

The three parts of this problem explore three approaches to answering these questions. In all three parts,

- the subscript i indicates treatment-specific quantities
- the subscript j indicates pig-specific quantities
- the subscript k indicates observation-specific quantities
- Y_{ijk}: weight of pig j in treatment i on week k
- X_{ijk}: week number (1, 2, or 3) for each observation.

1. a) Four possible models that might be fit to the data are shown below with the residual SS for each. Use this information to construct a test of H$_0$: $\beta_1 = \beta_2 = \beta_3$ against the alternative that the slopes are not all equal. You should compute an appropriate test statistic and indicate how you would compute the p-value. You do not need to actually find the p-value.

Number	Model	Residual SS
1	E $Y_{ijk} = \mu$	139,076.6
2	E $Y_{ijk} = \alpha + \beta X_{ijk}$	123,022.0
3	E $Y_{ijk} = \alpha_i + \beta X_{ijk}$	106,655.3
4	E $Y_{ijk} = \alpha_i + \beta_i X_{ijk}$	101,012.1

Answer: The hypothesis is the comparison between models 3 and 4. The residual SS can be used to compute an ANOVA table for that comparison. There are 45 observations (15 pigs * 3 obs/pig). Model 3 has 4 parameters (3 α's and 1 β). Model 4 has 6 parameters (3 α's and 3 β's). The SS for the hypothesis $\beta_1 = \beta_2 = \beta_3$ is R(4|3) = 106,655.3 - 101,012.1 = 5,643.2.

Hence:

Source	d.f.	SS	MS	F
Equal slopes	6-4 = 2	5,643.2	2,821.6	1.089
Residual	45-6 = 39	101,012.1	2,590.0	

b) Consider model 4, E $Y_{ijk} = \alpha_i + \beta_i X_{ijk}$. List the assumptions made by your test of $\beta_1 = \beta_2 = \beta_3$. Plotted below are a residual vs. predicted value plot and a normal quantile plot of the residuals. Use these diagnostics and the plot of the raw data to evaluate whether each of the assumptions is appropriate.

Answer:

Assumption:	Diagnostic	Evaluation
Normality	normal quantile plot	o.k.: Approx. straight line
Equal variance	residual plot	o.k.: Approx. equal spread
Independence	data plot	problem: lines for each pig are consistently above or below treatment average
Linearity	residual plot	o.k.: no visual curvature.

2. Another possible model is

$$Y_{ijk} = \alpha_{ij} + \beta_{ij} X_{ijk} + \varepsilon_{ijk},\ \varepsilon_{ijk} \sim \text{ iid N}(0, \sigma^2)$$

i.e. a separate regression line is fit to each animal. The parameters are estimated by least squares. The residual Mean Square for this model is 515.27, with 15 d.f. Estimates of the slope, $\hat{\beta}_{ij}$, for each pig in the no Vitamin E group and the high Vitamin E group are:

Treatment	Pig	$\hat{\beta}_{ij}$	s.e. $\hat{\beta}_{ij}$
None	1	-19.0	9.406
None	2	-47.5	9.406
None	3	-13.5	9.406
None	4	14.5	9.406
None	5	24.0	9.406
High	11	24.0	9.406
High	12	26.0	9.406
High	13	25.0	9.406
High	14	24.0	9.406
High	15	14.5	9.406

Define $\overline{\beta}_i$ as the mean slope of all individuals receiving the i'th treatment.

a) Estimate $\overline{\beta}_{\text{High}} - \overline{\beta}_{\text{None}}$.

Answer: A reasonable estimate of $\overline{\beta}_{\text{None}}$ is: (-19.0 + -47.5 -13.5 + 14.5 + 24)/5 = -8.3. A reasonable estimate of $\overline{\beta}_{\text{High}}$ is: (24.0 + 26.0 + 25.0 + 24.0 + 14.5)/5 = 22.7. Hence, a reasonable estimate of $\overline{\beta}_{\text{High}}$ -$\overline{\beta}_{\text{None}}$ is 22.7 - (-8.3) = 31.0.

---

b) Test the hypothesis that $\overline{\beta}_{\text{High}} = \overline{\beta}_{\text{None}}$. Calculate a test statistic and indicate how you would compute the p-value, but you do not need to report a p-value.

Answer: The estimate from part a) is a linear combination of the ten slopes, with coefficient vector **v** = [-0.2 -0.2 -0.2 -0.2 -0.2 0.2 0.2 0.2 0.2 0.2]'. The s.e. of a linear combination of the β_{ij} is s.e. $\beta_{ij}\sqrt{\mathbf{v'v}}$ = 9.406 $\sqrt{2/5}$ = 5.95. The estimate is normally distributed because it is a linear combination of the β_{ij}, which are linear combinations of the observations, which have normal distributions. The appropriate test statistic is the t statistic: 31.0/5.95 = 5.21. The p-value would be computed from the c.d.f of a 15 d.f. T distribution.

---

c) Suppose you only assumed that the ε_{ijk} were independent with E $\varepsilon_{ijk} = 0$ and Var $\varepsilon_{ijk} = \sigma^2$. Could you still justify your estimator in part a) and test procedure in part b)? Why or why not?

Answer: The estimator in part a) is still a reasonable estimator. The individual slopes are least squares estimates, so they are unbiased and minimum variance (among the class of linear unbiased estimators). The means of the slopes are linear combinations of the slopes, so the means are unbiased. Each slope has the same sampling variance, so the equally-weighted average is the minimum variance estimate.

The test procedure in part b) often reasonable in practice. If $X \sim N(0, \sigma^2), Vk/\sigma^2 \sim \chi^2_k$, and X and V independent, then $T = X/\sqrt{V/k}$ follows a t distribution. When the ε_{ijk} follow a normal distribution, all requirements for a t distribution are satisified. When the ε_{ijk}

are not assumed to have a normal distribution, $\overline{\beta}_{\text{High}} = \overline{\beta}_{\text{None}}$, is approximately normally distributed, because of independence of the ε_{ijk} and the Central limit theorem. The distribution of the sample variance and the correlation between the difference and the variance are unknown. Hence, there is no theoretical justification for the t distribution. However, theoretical and simulation studies have demonstrated that T-tests of differences are quite robust to non-normality.

---

3. A third possible model is the independent random coefficients model

$$Y_{ijk} = \alpha_{ij} + \beta_{ij} X_{ijk} + \varepsilon_{ijk}$$

$$\begin{aligned}
\varepsilon_{ijk} &\sim \text{iid N}(0, \sigma^2) \\
\begin{bmatrix} \alpha_{ij} \\ \beta_{ij} \end{bmatrix} &\sim \text{independent N}\left(\begin{bmatrix} \overline{\alpha}_i \\ \overline{\beta}_i \end{bmatrix}, \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix} \right) \\
\sigma^2 &\geq 0 \\
\sigma_a^2 &\geq 0 \\
\sigma_b^2 &\geq 0 \\
\text{Cov}\,(\alpha_{ij}, \varepsilon_{ijk}) &= 0 \\
\text{Cov}\,(\beta_{ij}, \varepsilon_{ijk}) &= 0
\end{aligned}$$

a) Assume that the three variance components (σ^2, σ_a^2, and σ_b^2) are known. This model can be rewritten as the model

$$Y_{ijk} = \overline{\alpha}_i + \overline{\beta}_i X_{ijk} + \gamma_{ijk},$$

which can be written in vector form as

$$Y = X\beta + \gamma$$

where the elements of γ are no longer independent. Define Σ as the variance-covariance matrix of γ. Express Σ in terms of the variance components, σ^2, σ_a^2, and σ_b^2.

Answer: Define the random effects: $\alpha_{ij}^* = \alpha_{ij} - \overline{\alpha}_i$ and $\beta_{ij}^* = \beta_{ij} - \overline{\beta}_i$ and $\mathbf{u} = [\alpha_{ij}^* \; \beta_{ij}^*]$. Then:

$$Y_{ijk} = \mathbf{X}\beta + \mathbf{X}\,\mathbf{u} + \varepsilon_{ijk}$$

So, $\gamma = X\, \boldsymbol{u} + \boldsymbol{\varepsilon}$, and $\Sigma = \text{E}\ \gamma\gamma'$. Hence,

$$\Sigma = \mathbf{X}\Psi\mathbf{X}' + \sigma^2\mathbf{I},$$

where $\Psi = \begin{bmatrix} \sigma_a^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$.

---

b) The elements of Σ are known when the three variance components, σ^2, σ_a^2, and σ_b^2, are known. In this case, what is a reasonable estimator for the vector of fixed effect parameters, β?

Answer: A reasonable estimator is the Generalized Least Squares estimate:

$$\hat{\beta} = (\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{X}'\Sigma^{-1}\mathbf{Y})$$

c) Consider estimating σ_b^2, the between-pig variance in slopes, using maximum likelihood. For these data (3 treatment groups), this model has 9 parameters: 3 mean intercepts, 3 mean slopes, and 3 variance components. Each variance component must be non-negative. One common approach to finding mle's is to set the score equations (the partial derivatives of the log likelihood with respect to each parameter) equal to zero, then solve that system of equations. In this model, does the solution to the score equations define the maximum likelihood estimate of the parameters? Why or why not?

Answer: Since there are constraints on the parameter space ($\sigma_b^2 \geq 0$), the m.l.e. is not necessarily the solution of the score equations. You must check two things:
1) that the solution to the score equations satisfies the constraint $\sigma_b^2 \geq 0$.
2) that there is not some point on the boundary of the parameter space with a larger log-likelihood.

d) Define $\hat{\sigma}_b^2$ as the mle of σ_b^2, the between-pig variance in slopes. Consider the likelihood ratio test of H$_0$: $\sigma_b^2 = \sigma_0^2$ vs. the alternative $\sigma_b^2 \neq \sigma_0^2$. For many choices of σ_0^2, the appropriate test statistic has an asymptotic Chi-square distribution with 1 d.f. The remaining parts of this question will evaluate whether this is the appropriate distribution for the likelihood ratio test of $\sigma_b^2 = 0$.

For simplicity, reduce the problem to a 1 parameter problem by conditioning on specific values of the other eight parameters. Define:

$l(x)$ as the log-likelihood function evaluated at $\sigma_b^2 = x$.
$D(x)$ as $-2(l(\sigma_0^2) - l(x))$.
σ_b^{2*} as the solution to the score equation, $\partial l/\partial\sigma_b^2 = 0$.
τ^2 as equal to $-1/\left.\frac{\partial^2 l(x)}{\partial x^2}\right|_{x=\sigma_b^{2*}}$.

First, we will imagine that there is no constraint on the parameter space for σ_b^2 and connect the asymptotic distribution of $D(x)$ to the asymptotic distribution of an unconstrained estimate of σ_b^2.

Use a Taylor series expansion of $l(\sigma_0^2)$ around σ_b^{2*} to relate $D(\sigma_b^{2*})$ to σ_b^{2*} and σ_0^2.

Answer: The second order Taylor series expansion of $l(x)$ around σ_b^{2*} is

$$l(x) \approx l(\sigma_b^{2*}) + (x - \sigma_b^{2*}) \left.\frac{\partial l(x)}{\partial x}\right|_{x=\sigma_b^{2*}} + \frac{1}{2}(x - \sigma_b^{2*})^2 \left.\frac{\partial^2 l(x)}{\partial x^2}\right|_{x=\sigma_b^{2*}} .$$

Since σ_b^{2*} is the solution to the score equations, $\left.\frac{\partial l(x)}{\partial x}\right|_{x=\sigma_b^{2*}} = 0$. Evaluating this at $x = \sigma_0^2$ gives

$$l(\sigma_0^2) \approx l(\sigma_b^{2*}) + \frac{1}{2}(\sigma_0^2 - \sigma_b^{2*})^2 \left.\frac{\partial^2 l(x)}{\partial x^2}\right|_{x=\sigma_b^{2*}} .$$

Hence,

$$\begin{aligned} D(\sigma_b^{2*}) &= -2(l(\sigma_0^2) - l(\sigma_b^{2*})) \\ &\approx -2(l(\sigma_b^{2*}) + \frac{1}{2}(\sigma_0^2 - \sigma_b^{2*})^2 \left.\frac{\partial^2 l(x)}{\partial x^2}\right|_{x=\sigma_b^{2*}} - l(\sigma_b^{2*})) \\ &= -(\sigma_0^2 - \sigma_b^{2*})^2 \left.\frac{\partial^2 l(x)}{\partial x^2}\right|_{x=\sigma_b^{2*}} , \text{ so:} \\ D(\sigma_b^{2*}) &\approx \frac{(\sigma_b^{2*} - \sigma_0^2)^2}{\tau^2} \end{aligned}$$

e) Relate the mle, $\hat{\sigma}_b^2$, to the unconstrained estimate, σ_b^{2*}, then use the approximation in d) to determine the asymptotic distribution of $D(\hat{\sigma}_b^2)$ under the null hypothesis, H$_0$: $\sigma_b^2 = 0$. Identify the appropriate critical value for an asymptotic α=0.05 test of $\sigma_b^2 = 0$ against the alternative: $\sigma_b^2 > 0$.

Answer:

Define $Z^* = \frac{(\sigma_b^{2*} - \sigma_0^2)}{\sqrt{\tau^2}}$. τ^2 is the negative inverse of the observed information, so τ^2 is a consistent estimate of the asymptotic variance of $\hat{\sigma}_b^2$. Under H$_0$, E $\hat{\sigma}_b^{2*} = \sigma_0^2 = 0$, so Z^* has an asymptotic standard normal distribution.

Define $f_M(x)$ as the p.d.f. of $\hat{\sigma}_b^2$, $f_U(x)$ as the p.d.f. of σ_b^{2*}, and $f_D(x)$ as the p.d.f. of $D(\hat{\sigma}_b^2)$, all under H$_0$: $\sigma_b^2 = 0$. Denote the p.d.f. of a 1 d.f. Chi-square distribution by $f_C(x)$.

$$\hat{\sigma}_b^2 = \begin{cases} 0 & \sigma_b^{2*} < 0 \\ \sigma_b^{2*} & \sigma_b^{2*} \geq 0 \end{cases} ,$$

so

$$f_M(x) = \begin{cases} P[\sigma_b^{2*} < 0] = 0.5 & x = 0 \\ f_U(x) & x > 0 \end{cases}$$

and

$$f_D(x) = \begin{cases} 0.5 & x = 0 \\ f_C(x)/2 & x > 0 \end{cases} ,$$

Hence, the c.d.f. of D($\hat{\sigma}_b^2$) is

$$F_D(x) = \begin{cases} 0.5 & x = 0 \\ 0.5 + F_C(x)/2 & x > 0 \end{cases},$$

where $F_C(x)$ is the c.d.f. of a 1 d.f. Chi-square distribution.

The critical value for an α=0.05 test of $\sigma_b^2 = 0$ is x such that $F_D(x) = 0.95$, i.e. $F_C(x) = 0.90$. The critical value is the 0.90 quantile of a 1 d.f. Chi-square distribution.