STATISTICS DEPARTMENT
M.S. EXAMINATION

PART II
OPEN BOOK

Tuesday, May 30, 2000

9:00 a.m. - 1:00 p.m.

Statistics Department Computer Lab, SC S152

Instructions: **Complete *four* of the five problems. Each problem counts 25 points. Unless otherwise noted, points are allocated approximately equally to lettered parts of a problem. Spend your time accordingly.**

Begin each problem on a new page. Write the problem number and the page number in the specified locations at the top of each page. Also write your chosen ID code number on every page. Please write only within the black borderlines, leaving at least 1" margins on both sides, top and bottom of each page. Write on one side of the page only.

At the end of this part of the exam you will turn in your answer sheets, but you will keep the question sheets and your scratch paper.

You may use a computer to work any of the problems, but your answers must be handwritten on standard paper provided for the examination. Printers may *not* be used during the exam, and pages printed out by computer may *not* be submitted. As indicated, some problems have data files available on disk.

1. The effectiveness ('effect', for short) of two medications (referred to below as sources 0 and 1) in combating a certain ailment is obtained for 3 different dose levels; the higher the 'effect' value, the better the outcome. The data is given below.

   (a) The variables 'source' and 'dose' are considered to be independent variables. Which of 'effect' or 'ln(effect)' would be a more appropriate dependent variable? Please support your choice. **NOTE: 'ln' denotes natural log.** use,

   (b) Suppose that we wished to compare the effectiveness of the two medications at various dose levels. Would these comparisons be easier if there was no interaction between 'source' and 'dose'? Please explain. Yes,

   (c) Is one of the two medications better than the other? If so, which is the better medication and by how much? Supply both an estimate and a confidence statement.

(The data below can be found at `I:\Courswrk\stat\Btrumbo\MSexam\medicate.mtw` **or** `medicate.dat` **or** `medicate.txt`**)**

Obs.	effect	source	dose
1	1.6274	0	1
2	1.5311	0	1
3	2.6434	0	1
4	1.0020	0	1
5	1.6608	0	1
6	1.6919	0	2
7	1.8477	0	2
8	3.6980	0	2
9	4.7505	0	2
10	2.9539	0	2
11	4.4478	0	3
12	4.5436	0	3
13	5.2340	0	3
14	4.8818	0	3
15	3.9323	0	3
16	3.7712	1	1
17	3.5277	1	1
18	5.5132	1	1
19	2.7331	1	1
20	5.3190	1	2
21	4.5851	1	2
22	4.8874	1	2
23	5.4799	1	2
24	11.2258	1	3
25	10.1708	1	3
26	9.9187	1	3
27	10.3526	1	3

2. Five laboratories each offer a service of testing the strength of fiberboard. The question is whether the results from the five laboratories are the same. To test this, two lots of fiberboard (two slightly different types) are selected by a manufacturer. From each lot 30 panels are selected, six of which are sent to each laboratory. Thus there are 2 x 5 x 6 = 30 strength measurements altogether, as shown in the table below. It is believed that measurements are normally distributed and that the inherent variability of measurements for each Lab x Lot combination is the same.

	Lab				
Lot	1	2	3	4	5
A	1483	1449	1499	1428	1509
	1496	1400	1472	1401	1439
	1441	1477	1483	1404	1416
	1416	1471	1509	1419	1441
	1450	1446	1489	1414	1419
	1478	1398	1435	1446	1444
B	1504	1465	1506	1407	1480
	1505	1423	1537	1416	1429
	1477	1418	1578	1455	1364
	1457	1445	1486	1435	1441
	1435	1424	1499	1423	1437
	1478	1426	1491	1442	1438

Data file (reading down columns) on I:\Courswrk\Stat\Btrumbo\MSexam\Fiblab.mtw, also Fiblab.txt.

(a) Write the most complete model supported by these data: use A or α (subscript i) for Lot and L or λ (subscript j) for Lab. For each factor use the Latin letter (A or L) if the effect is random and the Greek letter (α or λ) if the effect is fixed; briefly explain your choices. Show the ranges of the subscripts, any restrictions on parameters, and distributional assumptions about random variables in your model.

(b) Give the ANOVA table for this situation. At the 5% level of significance, which effects in your model are significant and which are not?

(c) If the Lab effect is significant, either use the Fisher LSD method to elaborate the pattern of differences in population means (if you think the effect is fixed), or estimate its variance component (if you think the effect is random). If the Lab effect is not significant, recommend how many strength measurements should have been made at each Lab in order to have a power of 90% in detecting the situation where one lab is giving measurements that average 20 points higher or lower than the other four labs.

(d) Perform specific tests to check whether the normality and homoscedasticity assumptions seem appropriate. Give the name of each test you choose, say whether the result is significant at the 5% level, and comment briefly on the consequences of your findings.

(e) We now reveal that the six measurements at each lab were made on three different days, two measurements each day. Thus for Lot A at Lab 1 the results 1483 and 1496 were obtained on one day of testing, 1441 and 1416 on another day, and 1450 and 1478 on still another day. (Each laboratory chooses randomly on which six days, scattered over several weeks, it will do the tests—three different days for each lot.) Because the testing equipment needs to be set up afresh each day, one wonders whether day-to-day differences within labs are an important source of variability in the measurements. Modify your model in (a) to accommodate this new information. Perform the ANOVA to see whether there is a significant Day effect (5% level).

close book #3

BT CB Sp 00

#3

Answers

3 (a) The cdf is $F(x) = 1 - \exp(-x/3), x > 0$. The pdf is $f(x) = (1/3) \exp(-x/3), x > 0$.

The mgf is $m(t) = \int_0^\infty \exp(tx)\exp(-x/3)dx = 1/(1-3t), t < 1/3$, where the integral is easily evaluated by combining the exponentials and making a change of variable.

[7 Points 2 for cdf, 2 for pdf, 3 for deriving mgf.]

(b) Mary's waiting time is the sum of four independent exponentials as in (a). Thus it has a gamma distribution with shape parameter 4 and scale parameter 3. The mean is 4(3) = 12; the variance is 4(9) = 36. The easiest derivation is to recognize that this is the distribution that has mgf $1/(1-3t)^4$. The previous time in service of the customer currently with the teller is irrelevant because of the no-memory property of the exponential distribution.

[6 Points 1 for correct gamma distribution, 1 for mean, 1 for variance, 2 for mgf argument, 1 for mention of no memory property]

(c) John's waiting time W to start service will be the *minimum* of four exponentials, which is an exponential with mean 3/4, and thus variance 9/16. $P(W > w) = [P(X > w)]^4 = [\exp(-w/3)]^4 = \exp(-4w/3)$, so that the cdf of W is $1 - \exp(-4w/3)$, which is the cdf of the claimed exponential. [In terms of reliability, this is like waiting for the failure of a series system of four components with identically distributed exponential lifetimes.]

[6 Points: 1 for correct exponential distribution, 1 for mean, 1 for variance, 3 for derivation of minimum.]

(d) The distribution is that of the *maximum* of four exponentials. Two approaches are possible:

First, one could argue that the waiting time for the first of the four to leave is exponential with mean 3/4 [as in (c)]; for the second, exponential with mean 3/3 = 1 because three of the original four remain; for the third, 3/2; and for the last, 3. Thus the mean of the maximum is 3/4 + 1 + 3/2 + 3 = 25/4 min. By independence (no memory), the variances also add: 9/16 + 1 + 9/4 + 9 = 205/16.

Second, one could derive the cdf of the maximum V as

$$F_V(v) = [F_X(v)]^4 = [1 - \exp(-v/3)]^4,\ v > 0,$$

differentiate to find the pdf, and use the pdf to derive the mean and variance.

[In terms of reliability, this is like waiting for the failure of a parallel system of four components with identically distributed exponential lifetimes.]

[6 Points: First way, 3 for mean, 3 for variance Second way, 2 for pdf, 2 for mean, 2 for variance]

3. The data in the following table were gathered for an environmental impact study that examined the relationship between the depth of an underground stream and the rate of its flow (Ryan, Joiner, and Ryan 1976). The data is available on the following webpage: http://www.telecom.csuhayward.edu/~esuess/data.htm

Depth	*FlowRate*
0.34	0.636
0.29	0.319
0.28	0.734
0.42	1.327
0.29	0.487
0.41	0.924
0.76	7.350
0.73	5.890
0.46	1.979
0.40	1.124

Use the appropriate SAS procedures to produce the computer output needed to answer the following questions. For each part that requires a plot you should describe what you see, no sketch or printout is required.

(a) Plot *FlowRate* versus *Depth* and describe the relationship between these two variables. Write down the SAS code used to produce the plot.

(b) Fit the linear regression model $FlowRate = \beta_0 + \beta_1 Depth + \epsilon$ to the data. Write down the SAS code used to fit the model. Give the estimated regression equation.

(c) Plot the residuals versus *Depth* and describe any problems with the linear regression model that are apparent from this plot.

(d) Transform both variables using the *log* transformation. Plot the transformed variables in a scatter plot. Do the data appear to be more linear after the transformation?

(e) Fit the linear regression model $log(FlowRate) = \beta_0 + \beta_1 log(Depth) + \epsilon$ to the data. Write down the SAS code used to fit the model. Give the estimated regression equation.

(f) Plot the residuals from this new model versus $log(Depth)$. Are there any signs of misfit?

(g) Which model fits the data better? Explain.

(h) An alternative approach to modeling this data is to use a linear model that includes a quadratic term. Fit the quadratic model $FlowRate = \beta_0 + \beta_1 Depth + \beta_2 Depth^2 + \epsilon$ to the data. Give the estimated regression equation.

(i) Of the three models which do you think is the best model for this data? Justify your answer.

4. Suppose that $X_1, X_2, ..., X_n$ are i.i.d. geometric random variables, where

$$p(x) = P(X = x) = (1-\theta)^{x-1}\theta, \quad x = 1, 2, ... \tag{1}$$

and $E[X_i] = 1/\theta$.

(a) Calculate the maximum likelihood estimate, $\hat{\theta}$, of θ.

(b) What is the asymptotic variance of the maximum likelihood estimator $\bar{\theta}$?

(c) Give an approximate $100(1-\alpha)\%$ confidence interval for θ based on the the maximum likelihood estimator $\bar{\theta}$.

(d) Derive the generalized likelihood ratio test of the null hypothesis $\theta = 0.5$ versus the alternative hypothesis $\theta \neq 0.5$. What is the form of the rejection region of the test? State the rejection region in terms of the likelihood ratio, Λ.

5. Five machines: M_1, M_2, M_3, M_4, and M_5 are used in a company's manufacturing process. The machines use a component that fails and must be replaced. A random sample of 10 of the components was obtained. The table below gives the times (in hours) until failure of each of two components in the sample, which were used in the machines:

Machine				
M_1	M_2	M_3	M_4	M_5
16, 20	22, 25	18, 21	32, 35	27, 28

Consider the model:

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, 2, 3, 4, 5;\ j = 1, 2$$

where y_{ij} is the time until failure of the *j*th component in M_i, and the ε_{ij} are independent random variables whose mean is zero and whose standard deviation is σ.

(a) Is $\mu + \alpha_1$ estimable? Prove your answer. If $\mu + \alpha_1$ is estimable, then find a 95% confidence interval for $\mu + \alpha_1$, assuming that the ε_{ij} have a normal distribution.

(b) Is α_1 estimable? Prove your answer. If α_1 is estimable, then find a 95% confidence interval for α_1, assuming that the ε_{ij} have a normal distribution.

(c) Is $\dfrac{\alpha_1 + \alpha_2 + \alpha_3}{3} - \dfrac{\alpha_4 + \alpha_5}{2}$ estimable? Prove your answer. If $\dfrac{\alpha_1 + \alpha_2 + \alpha_3}{3} - \dfrac{\alpha_4 + \alpha_5}{2}$ is estimable, then find a 95% confidence interval for $\dfrac{\alpha_1 + \alpha_2 + \alpha_3}{3} - \dfrac{\alpha_4 + \alpha_5}{2}$, assuming that the ε_{ij} have a normal distribution.

(d) Test whether the mean times until failure of the component differ among the machines at the 5% significance level. Assume that the ε_{ij} have a normal distribution.

Solution

~~Choose~~

(a) $\widehat{\ln}$ (effect). Since

	From Computer output	~~From effect (or ln effect)~~ versus dose	residuals ~~versus fitted~~ values
effect	lack of fit	Curvature especially at resource = 1	Curvature ~~in residual~~ plots
vs.			
ln (effect)	no lack of fit, higher r^2 (83.4% versus 80.4%)	Nice straight line for each resource	not so much over ~~here~~.

(b) Yes! Since then the comparison between the effectiveness of the two medications would depend on the dose level if there was interaction but not otherwise. Note from the computer output that there is no interaction.

(c) $E(\ln \text{effect}(\text{resource } 1)) - E(\ln \text{effect}(\text{resource } 2))$ (for any dose level)

$\approx .7560$ (see page 1 of computer output) $= \hat{\beta}_1$

an estimate of β_1 the coefficient of Source.

Can also get a 95% confidence interval for β_1

as $\beta_1 = .7560 \pm \underbrace{(2.0639)}_{\text{a } t \text{ with df}=24}(.1034)$ or $\beta_1 = .756 \pm .2134$

or $.5426 \le \beta_1 \le .9694$

Thus with 95% confidence, $E(\ln \text{effect}(\text{resource } 1)) - E\left(\ln \text{effect}\left(\text{resource } 2\right)\right) \ge .5426$.

Regression Analysis

The regression equation is
effect = - 1.43 + 3.36 source + 2.26 dose

Predictor	Coef	StDev	T	P
Constant	-1.4255	0.7002	-2.04	0.053
source	3.3606	0.5046	6.66	0.000
dose	2.2610	0.3071	7.36	0.000

S = 1.303 R-Sq = 80.4% R-Sq(adj) = 78.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	167.306	83.653	49.29	0.000
Residual Error	24	40.734	1.697		
Lack of Fit	3	26.192	8.731	12.61	0.000
Pure Error	21	14.541	0.692		
Total	26	208.040			

Source	DF	Seq SS
source	1	75.291
dose	1	92.015

Unusual Observations

Obs	source	effect	Fit	StDev Fit	Residual	St Resid
24	1.00	11.226	8.718	0.486	2.508	2.07R

R denotes an observation with a large standardized residual

Regression Analysis

The regression equation is
lneffect = - 0.026 + 0.756 source + 0.516 dose

Predictor	Coef	StDev	T	P
Constant	-0.0256	0.1435	-0.18	0.860
source	0.7560	0.1034	7.31	0.000
dose	0.51613	0.06291	8.20	0.000

S = 0.2669 R-Sq = 83.4% R-Sq(adj) = 82.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	8.6056	4.3028	60.40	0.000
Residual Error	24	1.7096	0.0712		
Lack of Fit	3	0.1231	0.0410	0.54	0.658
Pure Error	21	1.5866	0.0756		
Total	26	10.3152			

Source	DF	Seq SS
source	1	3.8105
dose	1	4.7951

Unusual Observations

Obs	source	lneffect	Fit	StDev Fit	Residual	St Resid
9	0.00	1.5582	1.0066	0.0689	0.5516	2.14R

R denotes an observation with a large standardized residual

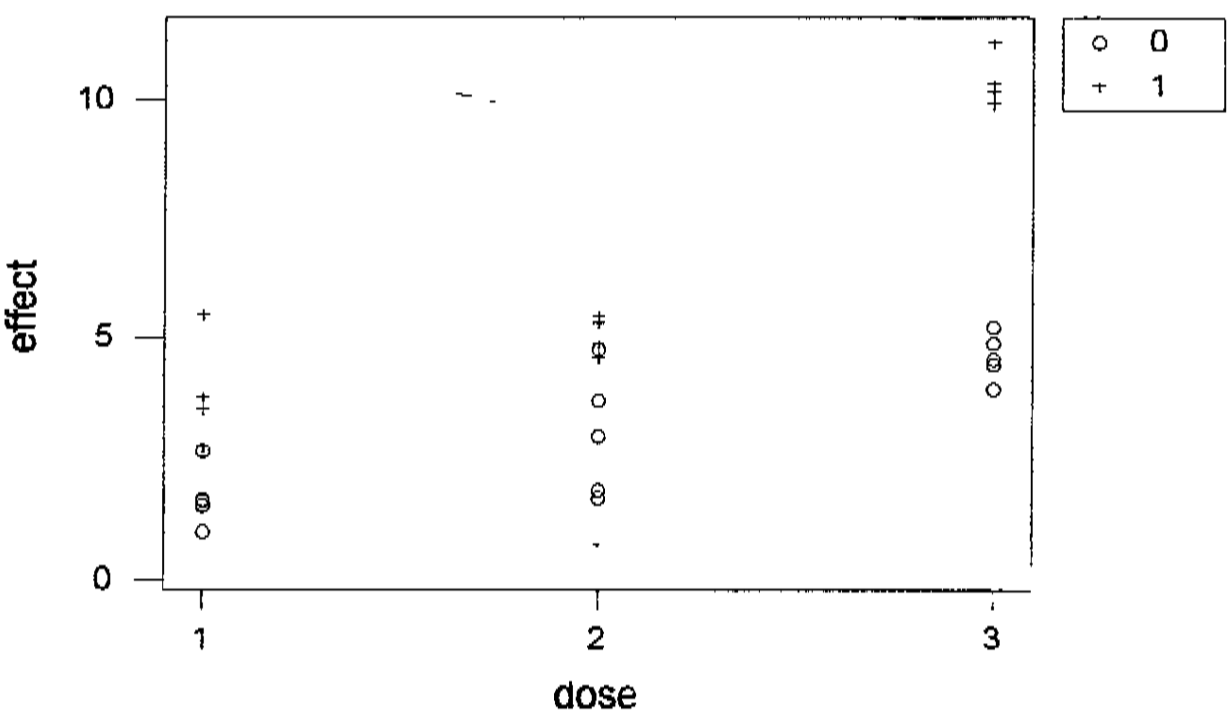
Regression Analysis

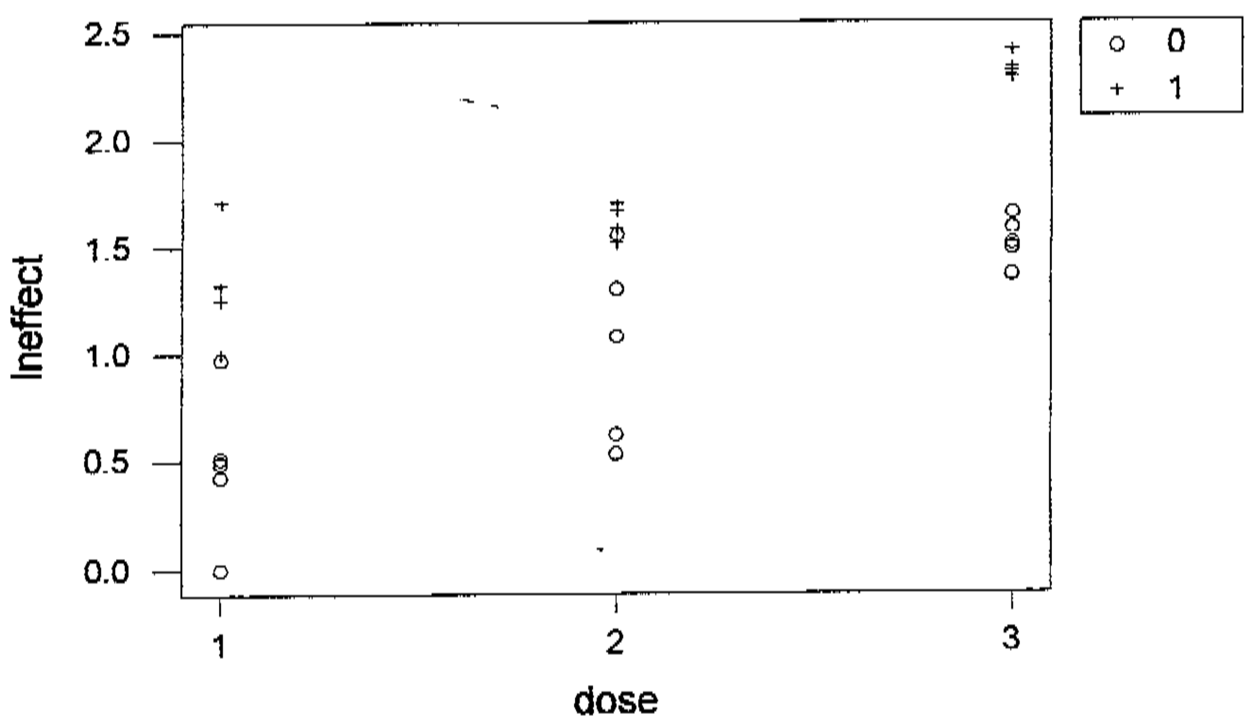
```
The regression equation is
lneffect = - 0.038 + 0.783 source + 0.522 dose - 0.014 sour*dos

Predictor        Coef       StDev          T        P
Constant      -0.0377      0.1862      -0.20    0.841
source         0.7832      0.2793       2.80    0.010
dose          0.52216     0.08620       6.06    0.000
sour*dos      -0.0136      0.1293      -0.10    0.917

S = 0.2726      R-Sq = 83.4%     R-Sq(adj) = 81.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3      8.6064      2.8688     38.61    0.000
Residual Error    23      1.7088      0.0743
Total             26     10.3152

Source       DF      Seq SS
source        1      3.8105
dose          1      4.7951
sour*dos      1      0.0008

Unusual Observations
Obs     source   lneffect         Fit   StDev Fit    Residual    St Resid
  9       0.00     1.5582      1.0066      0.0704      0.5516       2.09R

R denotes an observation with a large standardized residual
```

Data Display

Row	effect	source	dose
1	1.6274	0	1
2	1.5311	0	1
3	2.6434	0	1
4	1.0020	0	1
5	1.6608	0	1
6	1.6919	0	2
7	1.8477	0	2
8	3.6980	0	2
9	4.7505	0	2
10	2.9539	0	2
11	4.4478	0	3
12	4.5436	0	3
13	5.2340	0	3
14	4.8818	0	3
15	3.9323	0	3
16	3.7712	1	1
17	3.5277	1	1
18	5.5132	1	1
19	2.7331	1	1
20	5.3190	1	2

21	4.5851	1	2
22	4.8874	1	2
23	5.4799	1	2
24	11.2258	1	3
25	10.1708	1	3
26	9.9187	1	3
27	10.3526	1	3

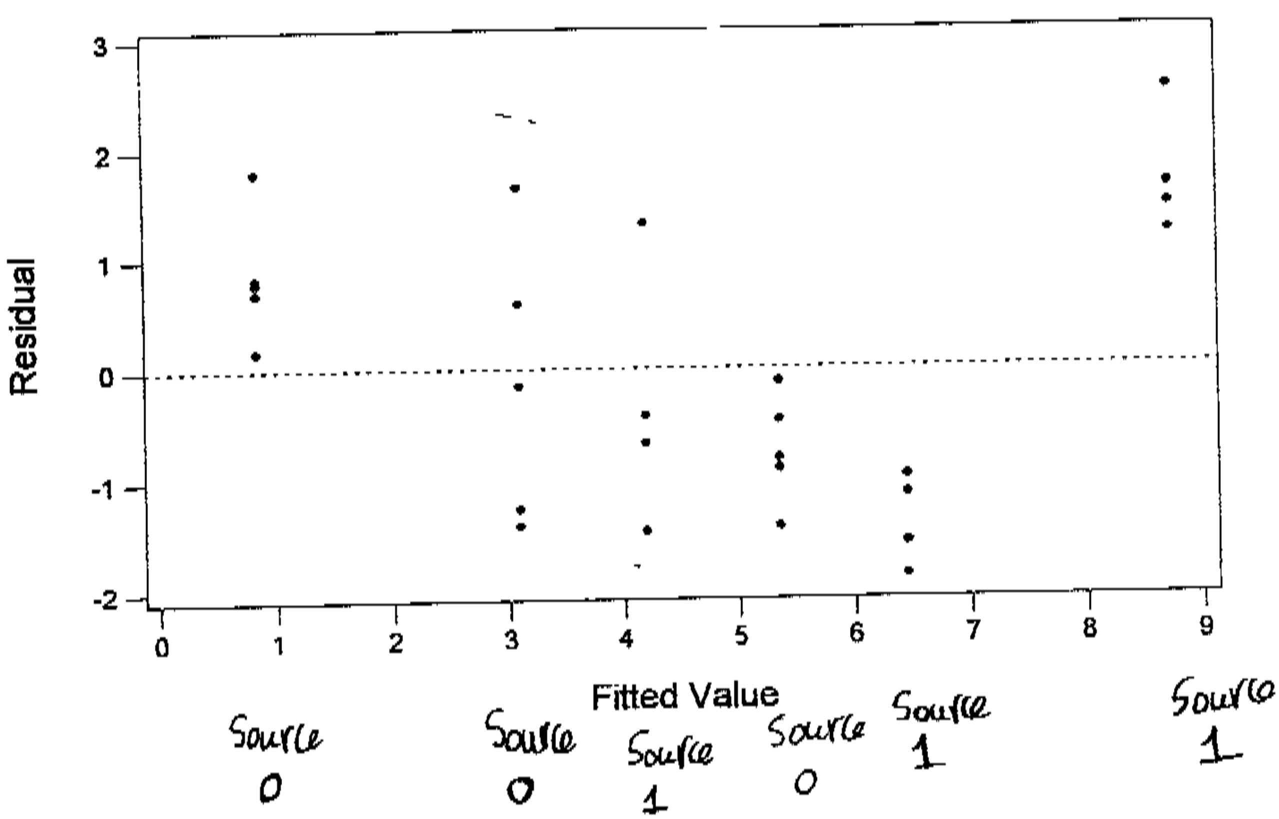

effect
10
5
0
1
2
3
dose
0
1

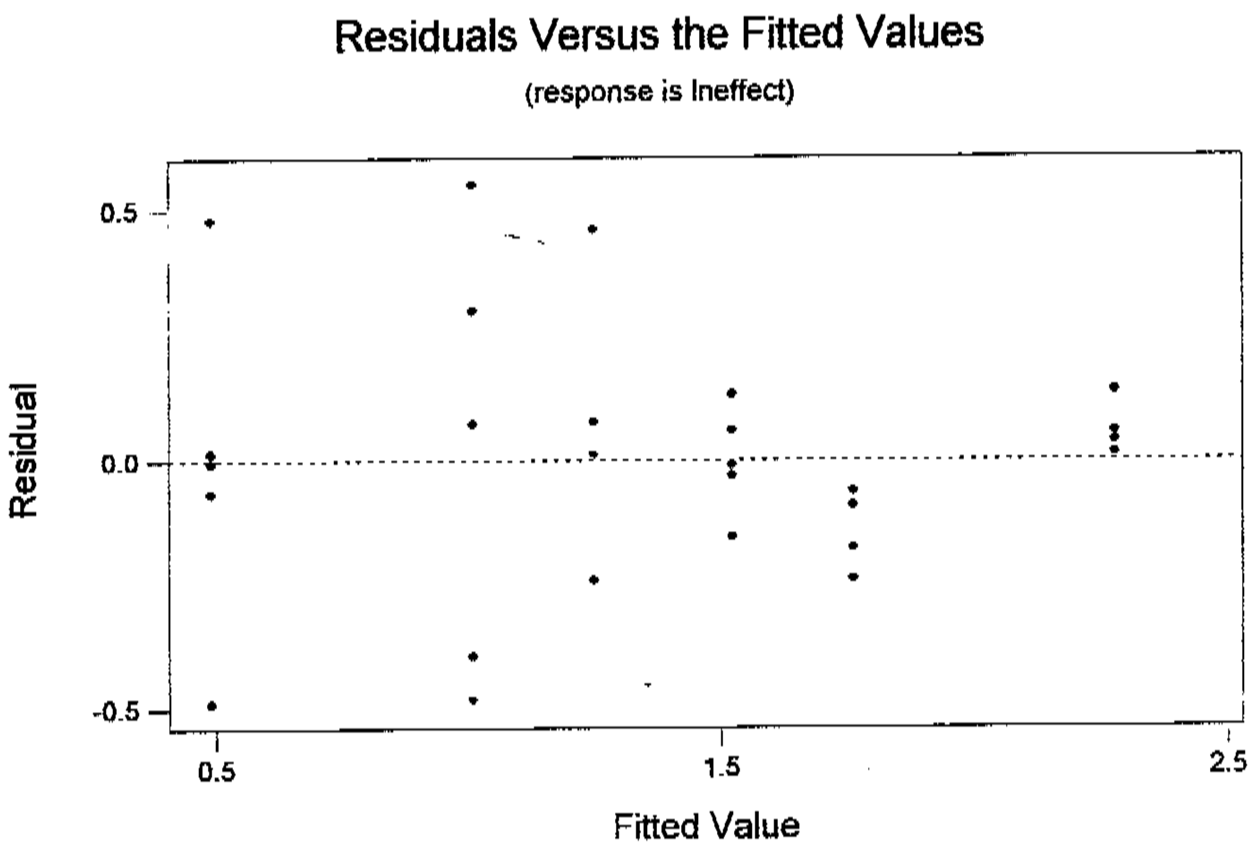


2.5
2.0
1.5
1.0
0.5
0.0
lneffect
1
2
3
dose
0
1

Residuals Versus the Fitted Values

(response is effect)





Residuals Versus the Fitted Values
(response is lneffect)
Residual
Fitted Value
0.5
0.0
-0.5
0.5
1.5
2.5

2.

(a) Both effects fixed, interaction supported.

$Y_{ijk} = \mu + \alpha_i + \lambda_j + (\alpha\lambda)_{ij} + e_{ijk}$

where $i = 1, 2$; $j = 1, 2, 3, 4, 5$; $k = 1, 2, 3, 4, 5, 6$.

$\Sigma_i\, \alpha_i = 0$, $\Sigma_j\, \lambda_j = 0$ $\Sigma_i\, (\alpha\lambda)_{ij} = 0$, $\Sigma_j\, (\alpha\lambda)_{ij} = 0$, e_{ijk} iid $N(0, \sigma^2)$.

(b)

```
MTB > set c3
DATA> 12(1:5)
DATA> end
MTB > anova Strength = Lot Lab Lot*Lab;
SUBC> restrict;
SUBC> ems;
SUBC> resids c4.
```

Analysis of Variance (Balanced Designs)

```
Factor      Type Levels Values
Lot        fixed      2      1      2
Lab        fixed      5      1      2      3      4      5

Analysis of Variance for Strength

Source        DF          SS          MS        F        P
Lot            1      1033.3      1033.3     1.26    0.266
Lab            4     43626.6     10906.7    13.33    0.000
Lot*Lab        4      4363.4      1090.8     1.33    0.271
Error         50     40912.8       818.3
Total         59     89936.2

Source        Variance Error Expected Mean Square for Each Term
              component term (using restricted model)
 1 Lot                    4    (4) + 30Q[1]
 2 Lab                    4    (4) + 12Q[2]
 3 Lot*Lab                4    (4) + 6Q[3]
 4 Error         818.3         (4)
```

No significant interaction. Lab effect very highly significant. Lot effect not significant.

(c) The Lab group means are shown below. Each is the average of 12 measurements.

Tabulated Statistics

```
Rows: Lab

   Strength
       Mean

1    1468.3
2    1436.8
3    1498.7
4    1424.2
5    1438.1
All  1453.2
```

$$\text{LSD} = t^* \sqrt{\frac{2\text{MSE}}{12}} = 2.01(13.48) = 27.1,$$

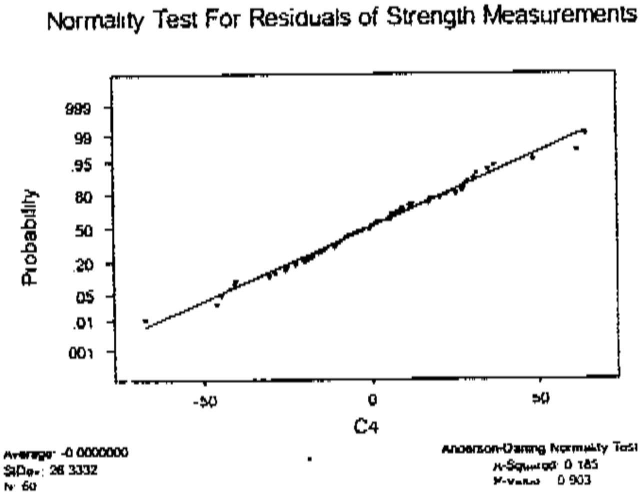
where $t^* = 2.01$ is the 0.025 point of $t(50)$, and MSE = 1090.8

Means in order:

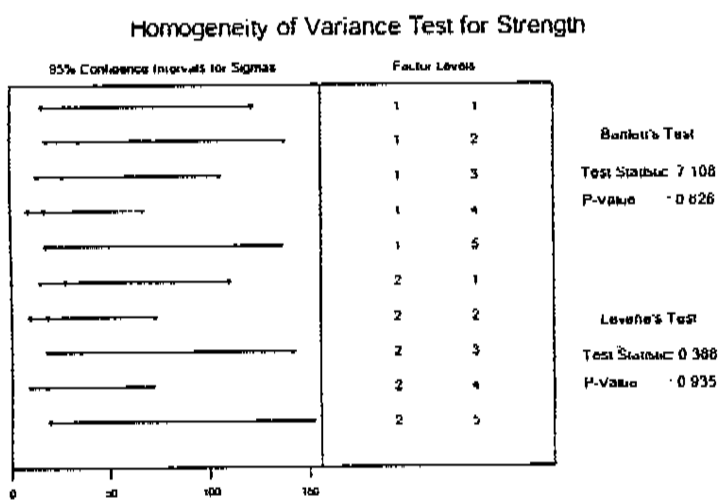
```
   4          2          5          2          3
----------------------------------------------------
1424.2     1436.8     1438.1     1468.3     1498.7
       12.6        1.3        30.2       30.4         Differences
```

So that the three Groups (4, 2, and 5) with the smallest sample means are not significantly different from each other. However, this cluster is significantly smaller than Group 2, which is significantly smaller than Group 3.

(d) Normal probability plot shows nearly a straight line. Anderson-Darling test does not reject normality.



Both Bartlett's and Levene's tests for homogeneity of variance fail to reject the null hypothesis of homogeneity. By treating this as a one-way ANOVA with 10 groups, one could use the F_{max} test as explained in Ott to obtain a similar result.



(e) Days nested within Lot x Lab cells. Introduce term $D(\alpha\beta)_{k(ij)}$ into the model iid $N(0, \sigma_D^2)$, $k = 1, 2, 3$. Change subscript on error variance to $l = 1,2$.

```
MTB > anova Strength = Lot | Lab Day(Lot Lab);
SUBC> random Day;
SUBC> restrict;
SUBC> ems.
```

Analysis of Variance (Balanced Designs)

Factor	Type	Levels	Values				
Lot	fixed	2	1	2			
Lab	fixed	5	1	2	3	4	5
Day(Lot Lab)	random	3	1	2	3		

Analysis of Variance for Strength

Source	DF	SS	MS	F	P
Lot	2-1= 1	1033.3	1033.3	1.05	0.318
Lab	5-1= 4	43626.6	10906.7	11.06	0.000
Lot*Lab	4	4363.4	1090.8	1.11	0.381
Day(Lot Lab)	20	19724.3	986.2	1.40	0.199
Error	30	21188.5	706.3		
Total	60-1 59	89936.2			

Source	Variance component	Error term	Expected Mean Square for Each Term (using restricted model)
1 Lot		4	(5) + 2(4) + 30Q[1]
2 Lab		4	(5) + 2(4) + 12Q[2]
3 Lot*Lab		4	(5) + 2(4) + 6Q[3]
4 Day(Lot Lab)	140.0	5	(5) + 2(4)
5 Error	706.3		(5)

The nested Day effect is not significant.

(#7)

3. Solution

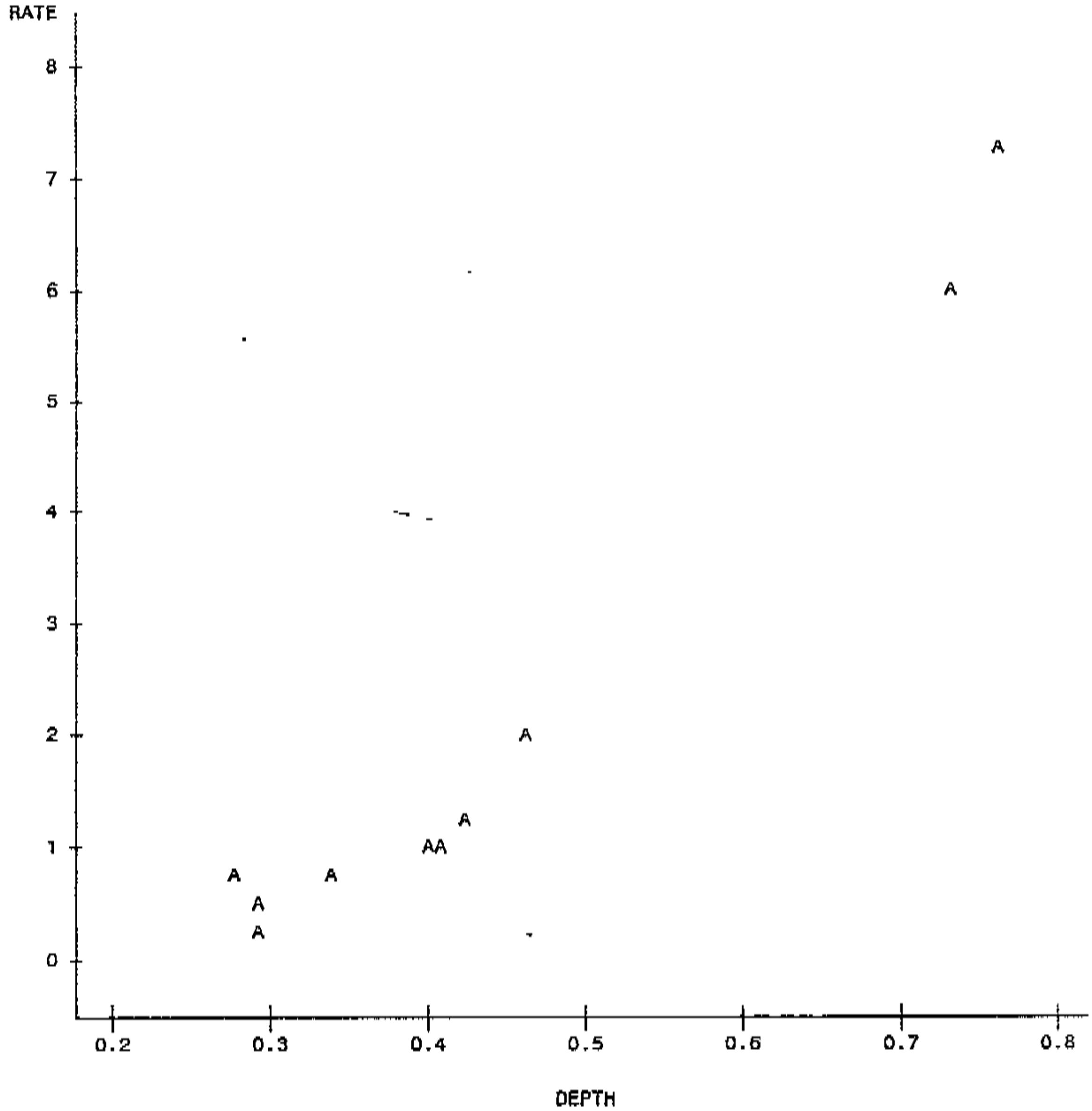
a.

The SAS code used to make the scatterplot:

```
*********************************************************************************************

DATA FLOW;
  INPUT DEPTH RATE;
  LOGDEPTH = LOG(DEPTH);
  LOGRATE = LOG(RATE);
  DEPTHSQ = DEPTH*DEPTH;
DATALINES;
  0.34 0.636
  0.29 0.319
  0.28 0.734
  0.42 1.327
  0.29 0.487
  0.41 0.924
  0.76 7.350
  0.73 5.890
  0.46 1.979
  0.40 1.124
;

PROC PLOT DATA=FLOW;
  PLOT RATE*DEPTH;
RUN;

*********************************************************************************************
```

Linear Relationship.

Plot of RATE*DEPTH. Legend: A = 1 obs, B = 2 obs, etc.



b.

The SAS code used to fit the linear regression model is:

```
*********************************************************************************

PROC REG DATA=FLOW;
  TITLE 'Regression for Rate versus Depth';
  MODEL RATE=DEPTH;
  PLOT RESIDUAL.*DEPTH;
RUN;

*********************************************************************************
```

Estimated regression equation:

FlowRate = -3.98 + 13.83 Depth

The CORR Procedure

2 Variables: RATE DEPTH

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
RATE	10	2.07700	2.46423	20.77000	0.31900	7.35000
DEPTH	10	0.43800	0.17332	4.38000	0.28000	0.76000

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	RATE	DEPTH
RATE	1.00000	0.97298 <.0001
DEPTH	0.97298 <.0001	1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: RATE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	51.73860	51.73860	142.07	<.0001
Error	8	2.91341	0.36418		
Corrected Total	9	54.65201			

Root MSE	0.60347	R-Square	0.9467
Dependent Mean	2.07700	Adj R-Sq	0.9400
Coeff Var	29.05490		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-3.98213	0.54298	-7.33	<.0001
DEPTH	1	13.83363	1.16061	11.92	<.0001

c.

For plot see next page.

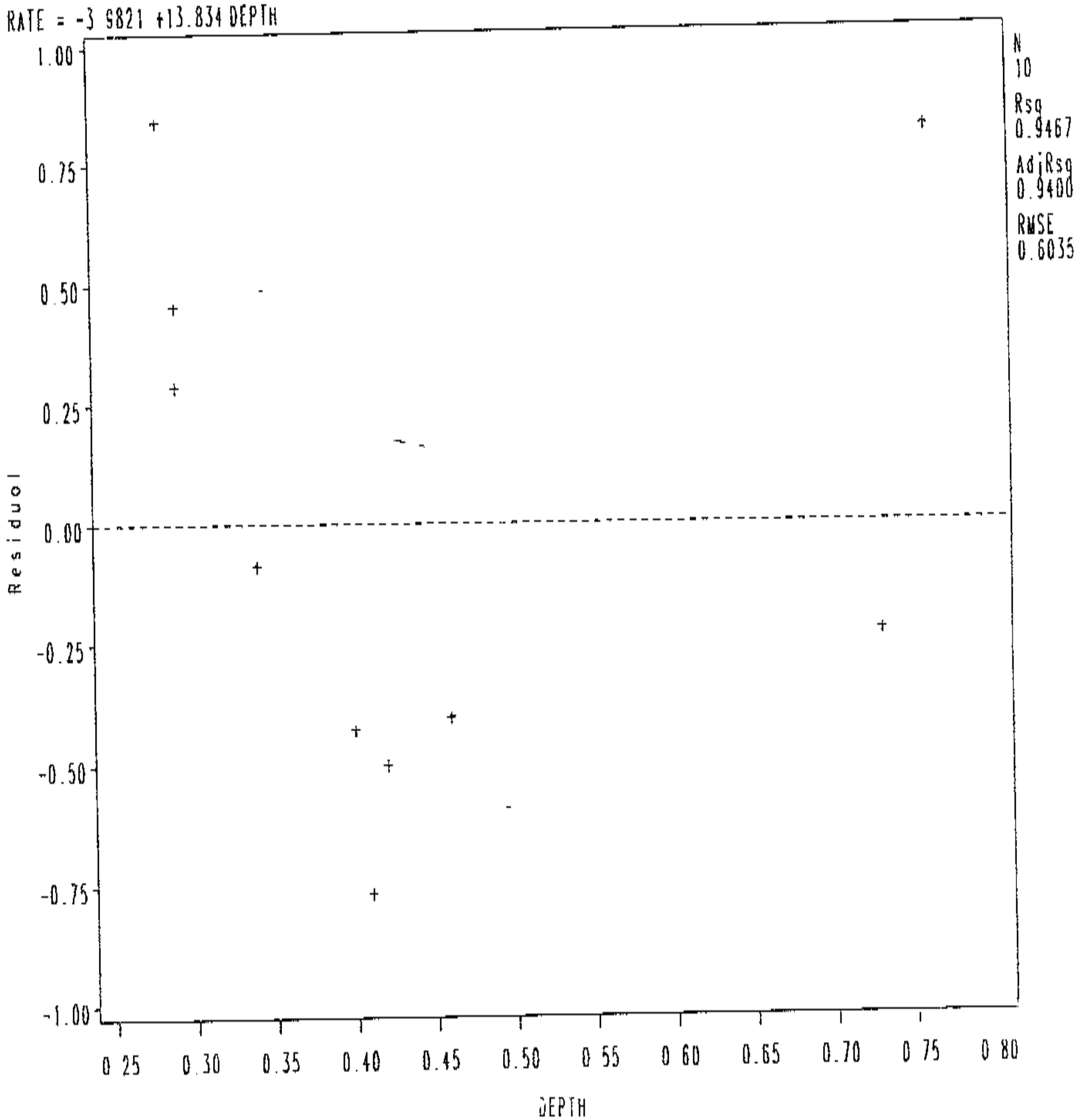
There is a problem with the residuals in that they do not appear to be random. There is a clear pattern in the residual plot.

d.

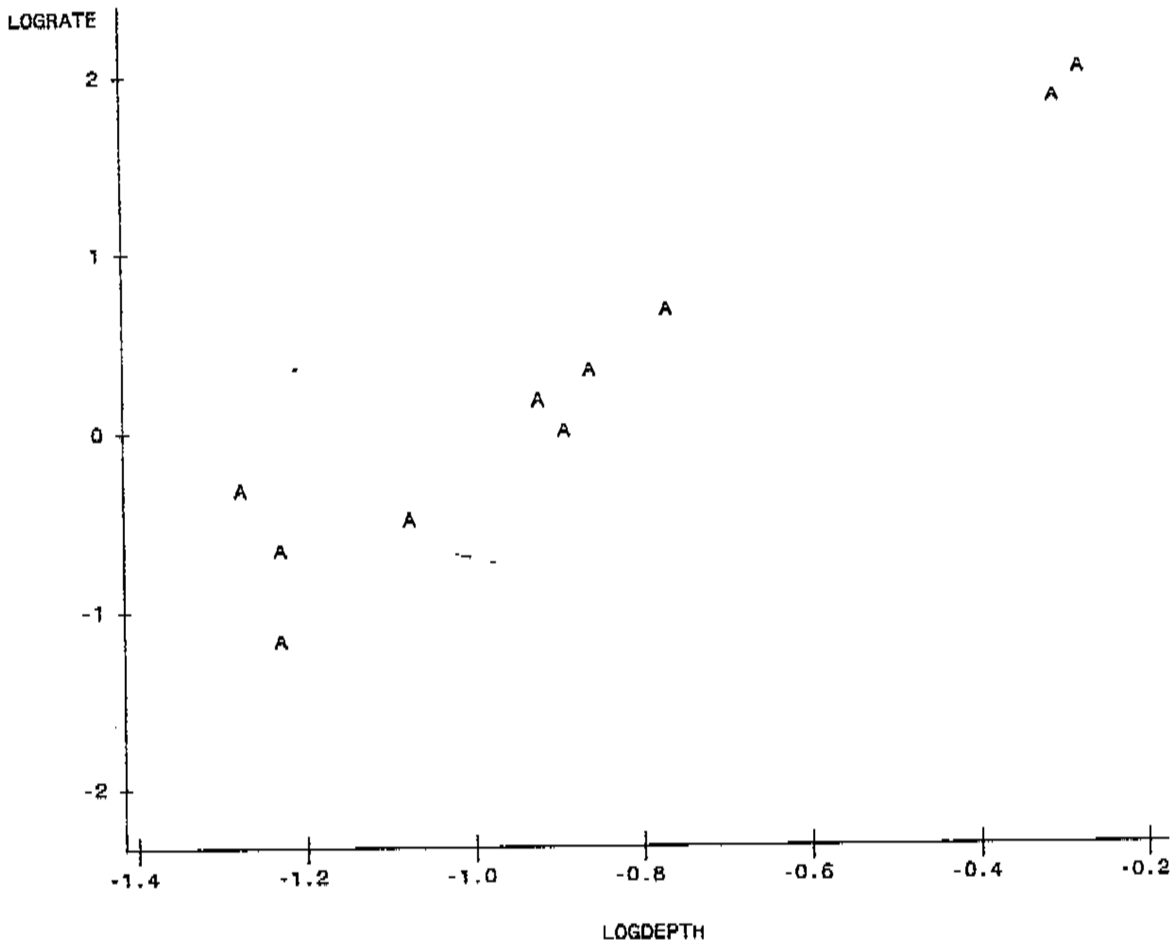
```
*************************************************************************************
PROC PLOT DATA=FLOW;
  PLOT LOGRATE*LOGDEPTH;
RUN;
*************************************************************************************
```

The data appear to be more linear after the transformation.

Regression for Rate versus Depth



Plot of LOGRATE*LOGDEPTH. Legend: A = 1 obs, B = 2 obs, etc.



e.

```
******************************************************************************

PROC REG DATA=FLOW;
  TITLE 'Regression for log(Rate) versus log(Depth)';
  MODEL LOGRATE=LOGDEPTH;
  PLOT RESIDUAL.*LOGDEPTH;
RUN;

******************************************************************************
```

log(FlowRate) = 2.66 + 2.76 log(Depth)

The CORR Procedure

2 Variables: LOGRATE LOGDEPTH

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
LOGRATE	10	0.21475	1.01934	2.14746	-1.14256	1.99470
LOGDEPTH	10	-0.88686	0.35718	-8.86859	-1.27297	-0.27444

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	LOGRATE	LOGDEPTH
LOGRATE	1.00000	0.96856 <.0001
LOGDEPTH	0.96856 <.0001	1.00000

The REG Procedure
Model: MODEL1
Dependent Variable: LOGRATE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.77270	8.77270	121.24	<.0001
Error	8	0.57886	0.07236		
Corrected Total	9	9.35156			

Root MSE	0.26899	R-Square	0.9381
Dependent Mean	0.21475	Adj R-Sq	0.9304
Coeff Var	125.26096		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	2.66614	0.23833	11.19	<.0001
LOGDEPTH	1	2.76413	0.25103	11.01	<.0001

f.

For plot see next page.

Looking at the residual plot there does not appear to be any misfit using this new model.

g.

There are two possible answers to this question: First, if the Rsq is used the untransformed model appears to fit better since Rsq = 0.9467 which is higher than the Rsq for the transformed model, Rsq = 0.9381. A second answer is to say that the transformed model fits the data better since the residual plot looks better.

h.

```
************************************************************************************

PROC REG DATA=FLOW;
  TITLE 'Regression for Rate versus Depth and Depth^2';
  MODEL RATE=DEPTH DEPTHSQ;
  PLOT RESIDUAL.*DEPTH;
RUN;

************************************************************************************
```

FlowRate = 1.68 - 10.86 Depth + 23.56 DepthSq

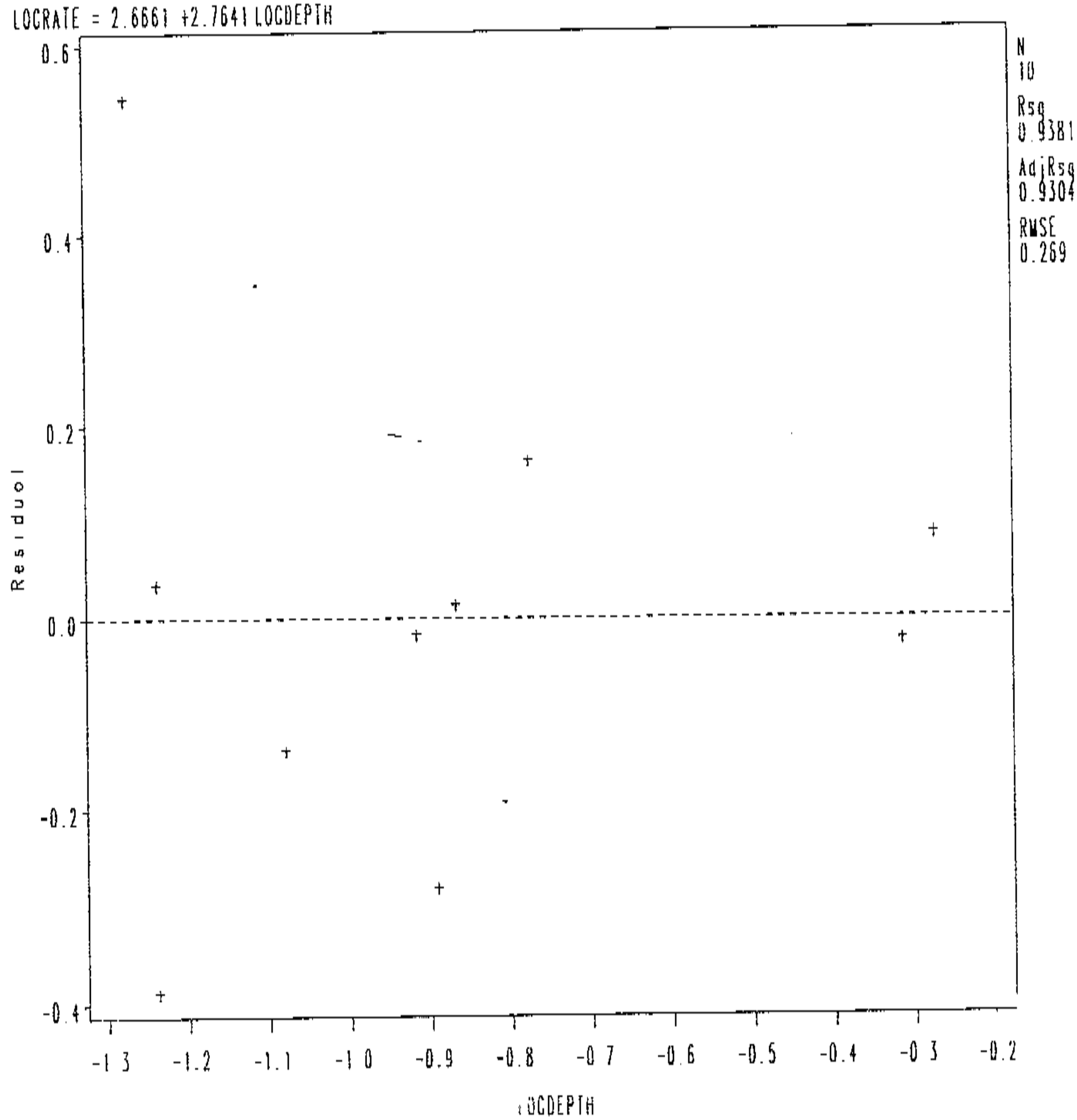
The REG Procedure
Model: MODEL1
Dependent Variable: RATE

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	54.10549	27.05275	346.50	<.0001
Error	7	0.54652	0.07807		
Corrected Total	9	54.65201			

Root MSE	0.27942	R-Square	0.9900
Dependent Mean	2.07700	Adj R-Sq	0.9871
Coeff Var	13.45294		

Regression for log(Rate) versus log(Depth)



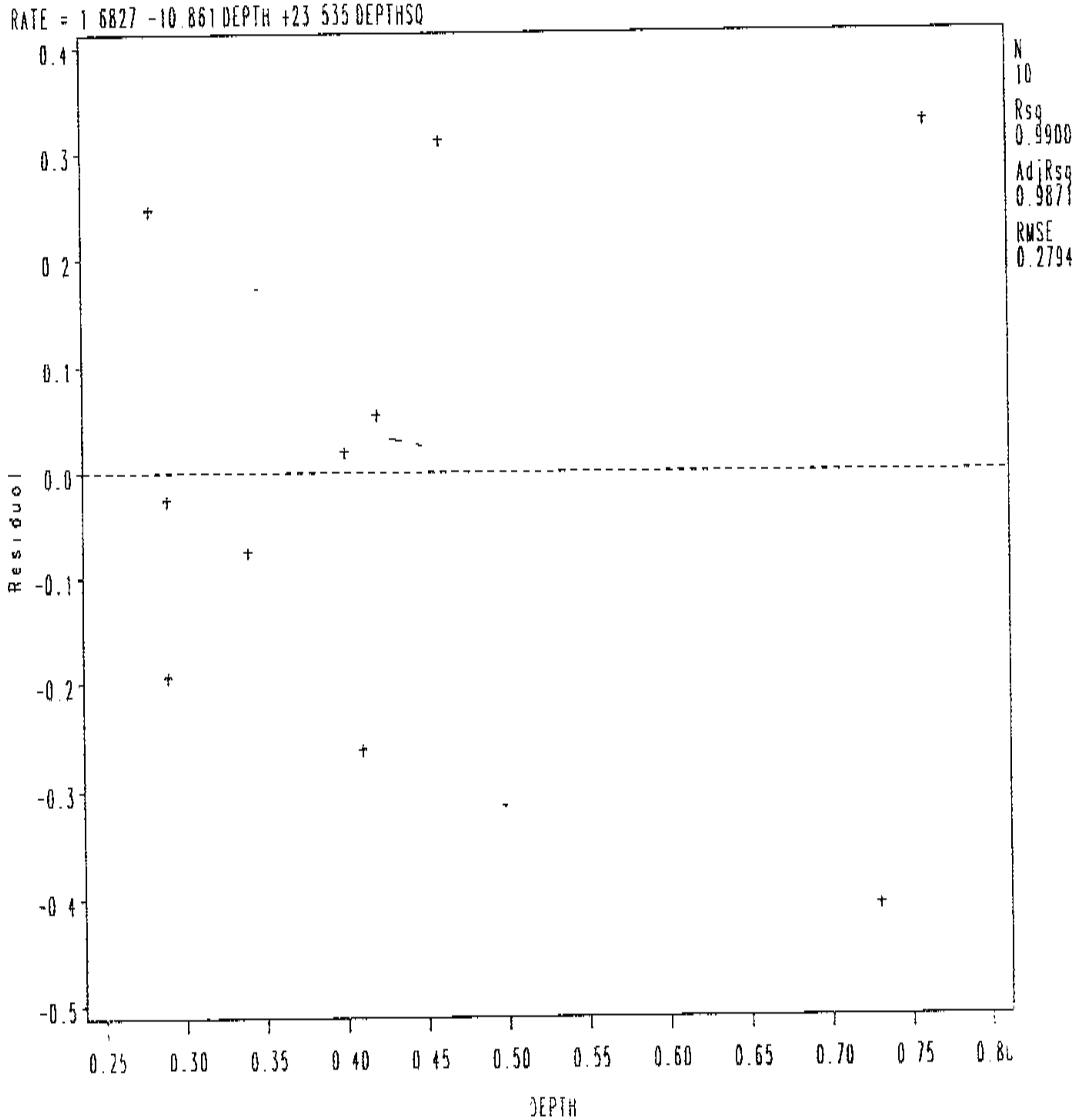
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1.68269	1.05912	1.59	0.1561
DEPTH	1	-10.86091	4.51711	-2.40	0.0472
DEPTHSQ	1	23.53522	4.27447	5.51	0.0009

1.

The last model that includes the squared term seems to fit the data the best since it has the most random looking residuals and the highest Rsq = 0.99.

Regression for Rate versus Depth and Depth^2



4. a) $L(\theta) = f(x_1, \dots, x_n | \theta) = \prod^n f(x_i|\theta) = \prod_{i=1}^n (1-\theta)^{x_i - 1}\theta$

$$= (1-\theta)^{\sum^n x_i - n} \theta^n = \left(\frac{\theta}{1-\theta}\right)^n (1-\theta)^{\sum^n x_i}$$

$$\ell(\theta) = n \log(\theta) - n \log(1-\theta) + \sum^n x_i \log(1-\theta)$$

$$\ell'(\theta) = \frac{n}{\theta} + \frac{n}{1-\theta} - \frac{\sum^n x_i}{1-\theta}$$

$\ell'(\theta) = 0$ so $\frac{n}{\theta} + \frac{n}{1-\theta} - \frac{\sum^n x_i}{1-\theta} = 0$

$$\frac{n \cancel{- n\theta} \cancel{+ n\theta}}{\theta\cancel{(1-\theta)}} = \frac{\sum^n x_i}{\cancel{1-\theta}}$$

$$\boxed{\hat{\theta} = \frac{1}{\bar{X}}}$$

b) $AV = \frac{1}{n I(\theta)}$

$$I(\theta) = E\left[\frac{\partial}{\partial\theta} \log f(x|\theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right]$$

$$f(x|\theta) = (1-\theta)^{x-1}\theta$$

$$\log f(x|\theta) = (x-1)\log(1-\theta) + \log(\theta)$$

$$\frac{\partial}{\partial\theta} \log f(x|\theta) = -\frac{x-1}{1-\theta} + \frac{1}{\theta} = -(x-1)(1-\theta)^{-1} + \theta^{-1}$$

$$\frac{\partial^2}{\partial\theta^2} \log f(x|\theta) = -\frac{x-1}{(1-\theta)^2} - \frac{1}{\theta^2}$$

Note: $E[X] = \frac{1}{\theta}$

$$I(\theta) = -E\left[-\frac{X-1}{(1-\theta)^2} - \frac{1}{\theta^2}\right] = \frac{1}{(1-\theta)^2} E[X-1] + \frac{1}{\theta^2}$$

$$= \frac{1}{(1-\theta)^2}\left(\frac{1}{\theta} - 1\right) + \frac{1}{\theta^2} = \frac{1}{(1-\theta)^2} \cdot \frac{1-\theta}{\theta} + \frac{1}{\theta^2}$$

$$= \frac{1}{\theta(1-\theta)} + \frac{1}{\theta^2} = \frac{\theta^2 + \theta - \theta^2}{\theta^3(1-\theta)} = \frac{1}{\theta^2(1-\theta)}$$

$$\therefore \boxed{AV = \frac{\theta^2(1-\theta)}{n}}$$

c) approximate $100(1-\alpha)\%$ CI for θ

since the MLE $\hat{\theta}$ has an asymptotic normal distribution with mean θ and $A.V. = \frac{\theta^2(1-\theta)}{n}$, an approximate $100(1-\alpha)\%$ CI for θ is

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}^2(1-\hat{\theta})}{n}}$$

$$\text{or. } \frac{1}{\bar{X}} \pm z_{\alpha/2}\sqrt{\frac{\frac{1}{\bar{X}^2}\left(1-\frac{1}{\bar{X}}\right)}{n}}$$

d) $H_0: \theta = 0.5$ vs. $H_1: \theta \neq 0.5$

G.L.R. $\Lambda = \dfrac{\max_{\Omega_0} L(\theta)}{\max_{\Omega} L(\theta)}$

$L(\theta) = \left(\dfrac{\theta}{1-\theta}\right)^n (1-\theta)^{\sum^n x_i}$ from a)

In the whole space Ω, the MLE of θ is $\hat{\theta} = ?$

so the denominator of Λ is

$$\max_{\Omega} L(\theta) = L(\hat{\theta}) = \left(\frac{\hat{\theta}}{1-\hat{\theta}}\right)^n (1-\hat{\theta})^{\sum^n x_i}$$

In the reduced space Ω_0. There is only a single point, i.e., simple hypothesis, so there is nothing to maximize

$$\max_{\Omega_0} L(\theta) = L(.5) = (.5)^{\sum x_i}$$

Thus $\Lambda = \dfrac{(.5)^{\sum x_i}}{\left(\dfrac{\hat{\theta}}{1-\hat{\theta}}\right)^n (1-\hat{\theta})^{\sum x_i}}$

Reject H_0 if $\Lambda < k$ s.t. $P(\Lambda < k) = \alpha$.

The large sample null distribution of $-2\log\Lambda$ is χ_1^2 since $df = \dim\Omega - \dim\Omega_0$

$$= 1 - 0 = 1$$

So the approximate form of the rejection region is

$$\text{Reject } H_0 \text{ when } -2\log\Lambda > \chi_1^2(\alpha).$$

5. (IX) $Y = X\beta + \varepsilon$ $\quad \beta' = (\mu, \alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5)$

$$X'Y = \begin{pmatrix} 244 \\ 36 \\ 47 \\ 39 \\ 67 \\ 55 \end{pmatrix}$$

$$X = \begin{pmatrix} 1_{10} & 1_2 & 0_2 & 0_2 & 0_2 & 0_2 \\ & 0_2 & 1_2 & & & 0_2 \\ & & & 1_2 & & \\ & & & & 1_2 & \\ & 0_2 & & & & 1_2 \end{pmatrix}$$

$$X'X = \begin{pmatrix} 10 & 2 & 2 & 2 & 2 & 2 \\ 2 & 2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 2 & & & \\ 2 & & & 2 & & \\ 2 & & & & 2 & 0 \\ 2 & 0 & & & 0 & 2 \end{pmatrix} \qquad X'X^c X'Y = \begin{pmatrix} 0 \\ 18 \\ 23.5 \\ 19.5 \\ 33.5 \\ 27.5 \end{pmatrix}$$

a conditional inverse of $X'X$ is

$$(X'X)^c = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 0 & & & & \\ 0 & & & & & 0 \\ 0 & & & & & \\ 0 & 0 & & & & 1/2 \end{pmatrix}$$

note: $t'\beta$ is estimable iff $* \; \exists \, c_{ij} \ni E\left(\sum_{ij} c_{ij} y_{ij}\right) = t'\beta$

) computations $\left(0 \; \tfrac{1}{3} \; \tfrac{1}{3} \; \tfrac{1}{3} \; -\tfrac{1}{2} \; -\tfrac{1}{2}\right)(X'X)^c \begin{pmatrix} 0 \\ \frac13 \\ \frac13 \\ \frac13 \\ -\frac12 \\ -\frac12 \end{pmatrix}$

$\left(0 \; \tfrac{1}{6} \; \tfrac{1}{6} \; \tfrac{1}{6} \; -\tfrac{1}{4} \; -\tfrac{1}{4}\right)$

$= 3/18 + \frac{2}{8} = \frac{24+36}{144} = \frac{60}{144} = \frac{5}{12}$

$\sqrt{MSE \; t'(X'X)^c t} = \sqrt{4.4 \cdot \left(\tfrac{5}{12}\right)} = 1.35400640$

then $$(X'X)^c (X'X) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(a) $\mu + \alpha_1 = (1 \; 1 \; 0 \; 0 \; 0 \; 0)\begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_5 \end{pmatrix}$, $\quad (1 \; 1 \; 0 \; 0 \; 0 \; 0)(X'X)^c X'X = (1 \; 1 \; 0 \; 0 \; 0 \; 0)$

or $E(y_{11}) = \mu + \alpha_1$

theorem from Myers + Milton hence $\mu + \alpha_1$ is estimable (MSE is next page)

95% CI $= (1 \; 1 \; 0 \; 0 \; 0 \; 0)(X'X)^c X'Y \pm t_{.025}\sqrt{MSE \; t'(X'X)^c t}$

$= 18 \pm 2.571\sqrt{4.4/2} = 18 \pm 3.81341$

(b) $\alpha_1 = (0 \; 1 \; 0 \; 0 \; 0)\begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_5 \end{pmatrix}$ $\quad t'(X'X)^c X'X = (1 \; 1 \; 0 \; 0 \; 0 \; 0) \neq t'$

1 $*$ $c_{i1} + c_{i2} = 0 \quad i = 2, 3, 4, 5$

requires $c_{11} + c_{12} = 0$ a contradiction; $c_{11} + c_{12} = 1$ $\quad \therefore \alpha_1$ is not estimable

(c) $\frac{\alpha_1 + \alpha_2 + \alpha_3}{3} - \frac{\alpha_4 + \alpha_5}{2} = \left(0 \; \tfrac{1}{3} \; \tfrac{1}{3} \; \tfrac{1}{3} \; -\tfrac{1}{2} \; -\tfrac{1}{2}\right)\begin{pmatrix} \mu \\ \alpha_1 \\ \vdots \\ \alpha_5 \end{pmatrix}$. $\quad t'(X'X)^c X'X = \left(0 \; \tfrac{1}{3} \; \tfrac{1}{3} \; \tfrac{1}{3} \; -\tfrac{1}{2} \; -\tfrac{1}{2}\right)$.

$\therefore \frac{\alpha_1 + \alpha_2 + \alpha_3}{3} - \frac{\alpha_4 + \alpha_5}{2}$ is est.

$E\left(\frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3}{3} - \frac{\bar{y}_4 + \bar{y}_5}{2}\right) = \nearrow$

95% $= \frac{18 + 23.5 + 19.5}{3} - \frac{33.5 + 27.5}{2} \pm 2.571\sqrt{MSE \; t'(X'X)^c t} = -10.17 \pm 3.481$

(d) Anova.

Source	SS	df	MS	F	$F_{.05}$
model	316.4	5-1=4	79.1	17.98 > 5.19	
error	22	10-5=5	4.4		
total	338.4	10-1=9			

$\sum_{ij} y_{ij} = 244$

$$SS_{tot} = 6292 - \frac{(244)^2}{10}$$

$\sum_{ij} y_{ij}^2 = 6292$

$$= 338.4$$

$SSE = SS_{tot} - SS_{model}$ $\quad SS_{model} = \frac{36^2}{2} + \frac{47^2}{2} + \frac{39^2}{2} + \frac{67^2}{2} + \frac{55^2}{2} - \frac{(244)^2}{10}$

$= 22$ $\quad = 316.4$

$H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5$

There is suff evidence that the means differ.

```
data machines;
input y x0 x1 x2 x3 x4 x5;
cards;
16 1 1 0 0 0 0
20 1 1 0 0 0 0
22 1 0 1 0 0 0
25 1 0 1 0 0 0
18 1 0 0 1 0 0
21 1 0 0 1 0 0
32 1 0 0 0 1 0
35 1 0 0 0 1 0
27 1 0 0 0 0 1
28 1 0 0 0 0 1
;
proc glm;
model y = x0 x1 x2 x3 x4 x5 /xpx i;
run;
proc glm;
model y = x0 x1 x2 x3 x4 x5 /noint xpx i;
estimate 'm1' x0 1 x1 1 x2 0 x3 0 x4 0 x5 0 ;
estimate 'm123 -m45' x1 .33333333 x2 .33333333 x3 .33333333 x4 -.5 x5 -.5;
run;
```

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	316.4000000	79.1000000	17.98	0.0036
Error	5	22.0000000	4.4000000		
Corrected Total	9	338.4000000			

Parameter	Estimate	Standard Error	t Value	Pr > |t|
m1	18.0000000	1.48323970	12.14	<.0001
m123 -m45	-10.1666666	1.35400639	-7.51	0.0007