STATISTICS DEPARTMENT
M.S. EXAMINATION

PART II
OPEN BOOK

Tuesday, May 21, 2002

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.**

The web site address for data and program files for this exam is:

http://www.sci.csuhayward.edu/~esuess/msexam/

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 variable that we are measuring is the proportion of inventory sold on a given day. We observe store 1 for a randomly selected 15 days and we observe store 2 for **another** randomly selected 15 days. Read data in **inventory** file from website.

day	store1	store2
1	0.1418	0.8500
2	0.2339	0.4445
3	0.5376	0.7238
4	0.4368	0.5526
5	0.4380	0.6748
6	0.4627	0.6640
7	0.5368	0.4831
8	0.4706	0.3571
9	0.5161	0.5746
10	0.3900	0.5313
11	0.2481	0.5087
12	0.5422	0.2816
13	0.5241	0.5030
14	0.3080	0.5012
15	0.5696	0.5111

1) Are each of the groups normally distributed? Support your answer.

2) Are the population variances of the two groups equal? Explain.

3) At the 3% significance level, can one conclude that the population median scores for the 2 stores are unequal? Answer this question 3 ways:

(a) By using the usual parametric t test.
(b) By ranking the data from 1 to 30 and using the formula in (a) on the ranks.
(c) By using an appropriate non-parametric test.

4) Which of the above approaches do you prefer? Please explain.

2. We compare 3 brands of a certain type of component (com.) in terms of population mean life-length (in months). Obtained are the life-lengths below for 11 randomly selected components of each brand. Read data in **component** file from website.

Com.	Brand A	Brand B	Brand C
1	52.2	46.7	75.2
2	56.4	60.5	63.7
3	57.1	58.9	73.2
4	46.9	62.9	66.2
5	49.1	65.8	67.4
6	52.5	53.3	69.4
7	63.0	66.9	70.4
8	52.0	70.9	72.3
9	61.1	73.7	63.6
10	55.3	65.8	61.9
11	46.2	70.2	74.4

(a) At the 5% significance level, can one conclude that there is a difference in life-length between the three brands? Please support your answer.

(b) If one is interested in comparing just brands B and C, at the 5% significance level, can one conclude that one of these two brands lasts longer? If so, with 95% confidence, at least how much longer does the better brand last?

(c) Suppose one is interested in comparing all pairs of brands, and wishes to be 95% confident about all conclusions that could be made about possible differences between these brands. Summarize your conclusions about the differences between these brands. If brands A and B are declared as being different, at least how much longer does the better of these two brands last?

(d) Suppose Brands A and C are from one manufacturer and Brand B is from another manufacturer. Can one conclude that the average of the Brand A and C population means is different from the Brand B population mean? Support your answer.

3.

Radial keratotomy reduces myopia for nearsighted patients by performing optical surgery in which radial incisions are made in the cornea. Cuts reduce the curvature of the cornea, thus improving vision for myopic patients.

Lynn[1] et al. (1987) examined the factors associated with five-year post-surgical change in error in vision. Measurements studied include error after five-years in diopters, baseline (prior to surgery) error in diopters, and baseline curvature also in diopters. (Myopic patients have negative errors. Patients who are far sighted have positive errors. Perfect vision has zero error.)

(a) Using a **SAS®** program, read in **the fabricated** file radial keratotomy.txt from the website. Columns are tab-delineated beginning with subject number, baseline error, baseline curvature, five-year post-operative error, diameter of clear zone (mm), patient gender, depth of incision scars (mm), baseline intraocular pressure (mm), and baseline central corneal thickness. Show the **SAS®** program including variable labels. (If you are unable to create this program, DO NOT STOP. CONTINUE WITH THIS PROBLEM. **SAS®** is required only in part (a).

(b) Using any software of your choice, run a regression using the data to predict five-year post-operative error from all available variables. Give the model and briefly comment on the fit.

(c) Create and include the interaction variable between gender and incision with the original variables. Perform an all-possible regressions analysis. Present the three best models and indicate the criterion measure that you are using and why you used it. Include the criterion value for these three best models.

(d) Find and report the best models using either a forward selection or a backward selection.

(e) What do the five models have in common? Discuss the benefits and problems with the models from parts (b), (c), and (d). Make a recommendation concerning radial keratotomy based on your results.

---

[1] Lynn M. J.; Waring, G. O., III; Sperduto, R. D.; et al. 1987. "Factors Affecting Outcome and Predictability of Radial Keratotomy in the PERK Study." *Archives of Opthalmology* 105: 42-31.

4. Consider the following method of estimating λ for the Poisson distribution. Observe that

$$p_0 = P(X = 0) = e^{-\lambda} \tag{1}$$

where $\lambda > 0$. Letting Y denote the number of zeros from an i.i.d. sample of n, λ might be estimated by

$$\tilde{\lambda} = -log\left(\frac{Y}{n}\right). \tag{2}$$

Note that $Y \sim Bin(n, p_0)$.

(a) Using the δ-method, obtain an approximate expression for the variance of this estimate of λ.

(b) Let $X_1, X_2, ..., X_n$ be a random sample from Poisson(λ). Compute the m.l.e. $\hat{\lambda}$ of λ. $= \frac{\sum X_i}{n}$

(c) Compute the exact variance $Var(\hat{\lambda})$ of $\hat{\lambda}$. $V(\hat{\lambda}) = V(\frac{\sum X_i}{n}) = \frac{1}{n^2} \cdot V(\sum X_i) = \frac{1}{n^2} \times n \times \lambda = \frac{\lambda}{n}$

(d) Compute the relative efficiency eff($\tilde{\lambda}, \hat{\lambda}$). Which estimator is more efficient? (Hint: Recall that $p_0 = e^{-\lambda}$.)

(e) Give a large sample $100(1 - \alpha)\%$ CI for λ using the m.l.e. $\hat{\lambda}$.

(f) Explain how you would use the parametric bootstrap to obtain an empirical 95% bootstrap CI for λ using $\tilde{\lambda}$. Also, explain how you would use the m.l.e. $\hat{\lambda}$.

(g) The Poisson distribution has been used by traffic engineers as a model for light traffic, based on the rational that if the rate is approximately constant and the traffic is light (so the individual cars move independently of each other), the distribution of counts in a given interval or space area should be nearly Poisson. The following table shows the number of right turns during 300 3-min intervals at an intersection. Run the following Splus program to implement the parametric bootstrap to produce the bootstrap confidence interval for λ using the two different estimators. Comment on the potential bias using $\tilde{\lambda}$ since $\hat{\lambda}$ is unbiased.

The Splus program pois.ssc is located at the following website:

http://www.sci.csuhayward.edu/~esuess/msexam/

n	0	1	2	3	4	5	6	7	8	9	10	11	12	13+
Frequency	14	30	36	68	43	43	30	14	10	6	4	1	1	0

```
### Splus program: pois.ssc

### relative efficiency of estimators of the Poisson rate lambda

# data

x <- c(rep(0,14),rep(1,30),rep(2,36),rep(3,68),rep(4,43),
       rep(5,43),rep(6,30),rep(7,14),rep(8,10),rep(9,6),
       rep(10,4),rep(11,1),rep(12,1))

brakes <- seq(0,15) - 0.5
hist(x, br = brakes)          # What does the distribution of the data
                              # look like?

n <- length(x)

# estimates of lambda

y <- 0                        # count the number of zeros
for(i in 1:n){
  if(x[i] == 0) y <- y + 1
}

lambda.tilde <- -log(y/n)  # first estimator
lambda.tilde
lambda.hat <- mean(x)         # m.l.e.
lambda.hat

# parametric bootstrap for lambda.tilde

B <- 1000  # number of bootstrap samples

lambda.tilde.star <- numeric(B)      # vector for storage
for(j in 1:B){
  x.boot <- rpois(n,lambda.tilde)    # bootstrap sample
  y <- 0
  for(i in 1:n){
    if(x.boot[i] == 0) y <- y + 1
  }
  if(y == 0) lambda.tilde.star[j] <- NA       # compute the estimator
    else lambda.tilde.star[j] <- -log(y/n)    # avoiding y = 0
}

# bootstrap analysis using lambda.tilde
```

```
mean(lambda.tilde.star)
sqrt(var(lambda.tilde.star))
quantile(lambda.tilde.star,c(0.025,0.975))

# parametric bootstrap for lambda.hat
B <- 1000  # number of bootstrap samples

lambda.hat.star <- numeric(B)          # vector for storage
for(j in 1:B){
  x.boot <- rpois(n,lambda.hat)        # bootstrap sample
  lambda.hat.star[j] <- mean(x.boot)  # compute the m.l.e.
}

# bootstrap analysis using lambda.hat

mean(lambda.hat.star)
sqrt(var(lambda.hat.star))
quantile(lambda.hat.star,c(0.025,0.975))

# plots for comparison

brakes <- seq(2,5,0.1)
par(mfrow=c(2,2))
hist(lambda.tilde.star, br = brakes)
hist(lambda.hat.star, br = brakes)

# estimated relative efficiency

eff <- var(lambda.hat.star)/var(lambda.tilde.star)
eff
```

5. A company tested 3 rods obtained from each of three vendors (V1, V2, V3). For each rod, the tensile strength (in suitable units) was recorded along with the level of a catalyst that the vendor used in making the rod.

		j=1	j=2	j=3
V1 (i=1)	Level of Catalyst x_{1j}	1	2	3
	Strength y_{1j}	5.2	4.4	2.1
V2 (i=2)	Level of Catalyst x_{2j}	1	2	3
	Strength y_{2j}	4.5	3.2	2.7
V3 (i=3)	Level of Catalyst x_{3j}	1	2	3
	Strength y_{3j}	3.7	2.4	1.5

(a) Ignore the catalyst and consider the model: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, $i = 1, 2, 3; j = 1, 2, 3$ where the ε_{ij}'s are independent; have mean 0 and standard deviation σ.

3 (i) Give the vector y, the vector β, and the matrix X used in expressing the above model as $y = X\beta + \varepsilon$.

3 (ii) Let $A = X'X$. Find a matrix A^c such that $AA^cA = A$. That is, find a conditional inverse of $X'X$.

In (iii) and (iv) below, show whether or not the indicated parameter is estimable:

4 (iii) τ_1

4 (iv) $\tau_1 - \tau_2$

6 (b) Consider the model $y_{ij} = \mu + \tau_i + \beta x_{ij} + \varepsilon_{ij}$, $i = 1, 2, 3; j = 1, 2, 3$.
where the ε_{ij}'s are independent, normal, have mean 0 and standard deviation σ.

For this model, a computer package produces the following ANOVA:

```
Source                    DF    Sum of Squares    Mean Square
Model                      4       109.3383333     27.3345833
Error                      5         0.9516667      0.1903333
Uncorrected Total          9       110.2900000
```

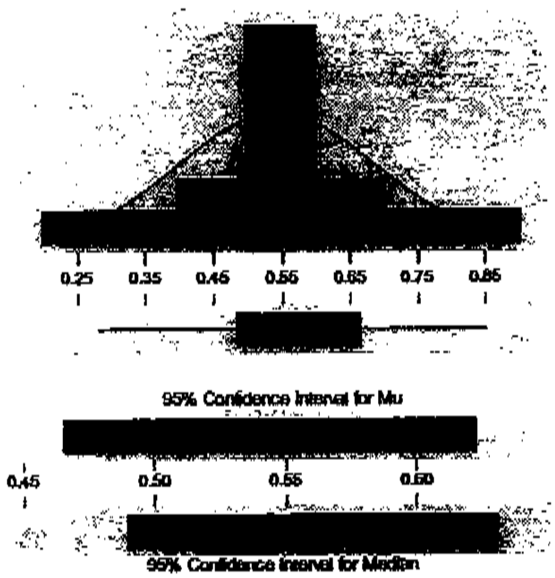
For the model $y_{ij} = \mu + \beta x_{ij} + \varepsilon_{ij}$, $i = 1, 2, 3; j = 1, 2, 3$, the computer produces the following ANOVA:

```
Scource                   DF    Sum of Squares    Mean Square
Model                      2       106.4116667     53.2058333
Error                      7         3.8783333      0.5540476
Uncorrected Total          9       110.2900000
```

Using the above results, test the null hypothesis H_0: $\tau_1 = \tau_2 = \tau_3$ at the 5% significance level.

Answer #1 OB

Descriptive Statistics



Variable: store2

Anderson-Darling Normality Test

A-Squared:	0.377
P-Value:	0.362
Mean	0.544093
StDev	0.141885
Variance	2.01E-02
Skewness	0.356143
Kurtosis	0.673897
N	15
Minimum	0.281600
1st Quartile	0.483100
Median	0.511100
3rd Quartile	0.664000
Maximum	0.850000

95% Confidence Interval for Mu

0.465514	0.622672

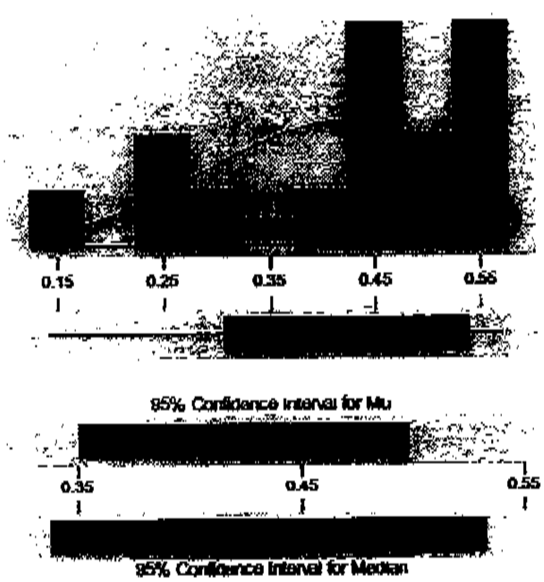
95% Confidence Interval for Sigma

0.103885	0.223783

95% Confidence Interval for Median

0.489850	0.630609

Descriptive Statistics



Variable: store1

Anderson-Darling Normality Test

A-Squared:	0.730
P-Value:	0.045
Mean	0.423753
StDev	0.132195
Variance	1.75E-02
Skewness	-8.5E-01
Kurtosis	-2.0E-01
N	15
Minimum	0.141800
1st Quartile	0.308000
Median	0.462700
3rd Quartile	0.535800
Maximum	0.589600

95% Confidence Interval for Mu

0.350546	0.496961

95% Confidence Interval for Sigma

0.096784	0.208485

95% Confidence Interval for Median

0.338627	0.532056

Descriptive Statistics: store1, store2

Variable	N	Mean	Median	TrMean	StDev	SE Mean
store1	15	0.4238	0.4627	0.4342	0.1322	0.0341
store2	15	0.5441	0.5111	0.5408	0.1419	0.0366

Variable	Minimum	Maximum	Q1	Q3
store1	0.1418	0.5696	0.3080	0.5368
store2	0.2816	0.8500	0.4831	0.6640

Test for Equal Variances: store1 vs store2

F-Test (normal distribution)

Test Statistic: 0.868
P-Value : 0.796

Levene's Test (any continuous distribution)

Test Statistic: 0.000
P-Value : 0.987

Two-Sample T-Test and CI: store1, store2

Two-sample T for store1 vs store2

	N	Mean	StDev	SE Mean
store1	15	0.424	0.132	0.034
store2	15	0.544	0.142	0.037

Difference = mu store1 - mu store2
Estimate for difference: -0.1203
95% CI for difference: (-0.2229, -0.0178)
T-Test of difference = 0 (vs not =): T-Value = -2.40 P-Value = 0.023 DF = 28
Both use Pooled StDev = 0.137

Two-Sample T-Test and CI: rank1, rank2

```
Two-sample T for rank1 vs rank2

        N      Mean     StDev   SE Mean
rank1  15     12.40      8.26       2.1
rank2  15     18.60      8.47       2.2

Difference = mu rank1 - mu rank2
Estimate for difference:  -6.20
95% CI for difference: (-12.46, 0.06)
T-Test of difference = 0 (vs not =): T-Value = -2.03  P-Value = 0.052  DF = 28
Both use Pooled StDev = 8.36
```

Mann-Whitney Test and CI: store1, store2

```
store1     N =  15     Median =      0.4627
store2     N =  15     Median =      0.5111
Point estimate for ETA1-ETA2 is     -0.0945
95.4 Percent CI for ETA1-ETA2 is (-0.2260,0.0054)
W = 186.0
Test of ETA1 = ETA2  vs  ETA1 not = ETA2 is significant at 0.0564

Cannot reject at alpha = 0.05
```

<u>Summary</u>

1) Store 1 data is not normal; store 2 data looks better.
2) Equality of variances can be assumed.
3) We have significance at the 3% level with the usual parametric test but not with the other two tests.
4) Because of the lack of normality go with the latter two tests which are pretty comparable.

Answer # 2 Q3

One-way ANOVA: life versus brand

```
Analysis of Variance for life
Source     DF        SS          MS         F          P
brand       2    1369.9       685.0     18.96      0.000
Error      33    1192.2        36.1
Total      35    2562.1
                                    Individual 95% CIs For Mean
                                    Based on Pooled StDev
Level       N      Mean       StDev  ------+---------+---------+---------+
1          12    54.083       5.252  (-----*-----)
2          12    63.708       7.812            (-----*-----)
3          12    68.983       4.447                   (-----*-----)
                                     ------+---------+---------+---------+
Pooled StDev =    6.011                  54.0      60.0      66.0      72.0

Tukey's pairwise comparisons

    Family error rate = 0.0500
Individual error rate = 0.0196

Critical value = 3.47

Intervals for (column level mean) - (row level mean)

                 1           2

       2   -15.646
            -3.604

       3   -20.921     -11.296
            -8.879       0.746


Fisher's pairwise comparisons

    Family error rate = 0.120
Individual error rate = 0.0500

Critical value = 2.035

Intervals for (column level mean) - (row level mean)

                 1           2

       2   -14.618
            -4.632

       3   -19.893     -10.268
            -9.907      -0.282
```

Contrast Coefficients

	BRAND		
Contrast	1.00	2.00	3.00
1	1	-2	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
LIFE	Assume equal variances	1	-4.3500	4.2501	-1.024	33	.314
	Does not assume equal	1	-4.3500	4.9283	-.883	15.385	.391

summary

1) Yes! See the above ANOVA.
2) From Fisher's test, there is a difference and the better brand, brand C, lasts, on the average, at least .282 months longer.
3) From Tukey's test, Brands B and C are both better than Brand A. Brand B lasts at least 3.604 months longer than Brand A.
4) Can't conclude difference. See contrast material above.

..nswer #3 01S

Solution Regression Model for Radial Keratotomy:

a. Program part A: This program will fulfill all portions of the problem except part (e.) Only the lines up to the first Proc print are required for part A.

```
DATA KERA;
INFILE "C:\My Documents\department\masters\exam questions
 spring 2002\radial keratotomy3.txt"
                 DELIMITER='09'x;
INPUT VAR1-VAR5 VAR6 $ VAR7-VAR9;
LABEL VAR1='SUBJECT'
      VAR2='BASE ERROR'
      VAR3='BASE CURVE'
      VAR4='POST ERROR'
      VAR5='CLEAR ZONE'
      VAR6='SUBJECT GENDER'
      VAR7='INCISION'
      VAR8='BASE PRESSURE'
      VAR9='THICKNESS'
      NGEN='GENDER 1 m 0 f'
      VAR10='INTER GENDER INCISION';
NGEN=1;
IF VAR6='F' THEN NGEN=0;
VAR10=NGEN*VAR7;
* The preceding lines satisfy part a of the regression question
        and creates the interaction variable gender by incision;
PROC PRINT; RUN;
PROC CORR; RUN;
*The following lines satisfy part b of the regression question,
      including all variables except the interaction term in the
model;
PROC REG;
MODEL VAR4=VAR1 - VAR3 VAR5 VAR7 -- NGEN/P R;
*The following lines satisfy part c of the regression question,
      performing an all variables regression using adjusted rsquare as
criterion;
proc reg data=kera outest=modelsr;
model var4=var1 - var3 var5 var7 -- var10/selection=adjrsq rsquare cp;
title 'Best models by adjusted rsquare';
proc print data=modelsr;
proc reg data=kera outest=modelsc;
model var4=var1 - var3 var5 var7 -- var10/selection=cp adjrsq rsquare;
title 'Best models by Mallows cp statistic';
proc print data=modelsc;
*The following lines satisfy part d of the regression question,
      performing both forward and backward regression;
proc reg data=kera;
model var4=var1 - var3 var5 var7 -- var10/selection=backward;
title 'Best model according to backward selection';
proc reg data=kera;
model var4=var1 - var3 var5 var7 -- var10/selection=forward;
title 'Best model according to forward selection';
RUN:
```

b. All terms in model.

Even though the R-squared value is only 33%, residual lots do not indicate outliers or obvious unusual points. Other terms such as interactions or powers of the independent variable might be tried, but no really clear indications are present. Subject is a silly variable to include and is not significant. Only the variables base curve and clear zone are significant in this model. Some pruning of variables should be considered. All output is not necessary. Discussion of R-squared and terms in the model that seem important.

The REG Procedure
Model: MODEL1
Dependent Variable: VAR4 POST ERROR

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	26.51084	3.31386	2.77	0.0139
Error	45	53.74210	1.19427		
Corrected Total	53	80.25294			

Root MSE	1.09283	R-Square	0.3303
Dependent Mean	3.83343	Adj R-Sq	0.2113
Coeff Var	28.50782		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	18.32156	5.96371	3.07	0.0036
VAR1	SUBJECT	1	-0.00672	0.01089	-0.62	0.5404
VAR2	BASE ERROR	1	0.13180	0.60324	0.22	0.8280
VAR3	BASE CURVE	1	-0.29626	0.12766	-2.32	0.0251
VAR5	CLEAR ZONE	1	-0.64544	0.20889	-3.09	0.0034
VAR7	INCISION	1	0.35699	0.50113	0.71	0.4799
VAR8	BASE PRESSURE	1	-0.27749	0.63433	-0.44	0.6639
VAR9	THICKNESS	1	0.37258	0.55218	0.67	0.5033
NGEN	GENDER 1 m 0 f	1	-1.34449	1.00140	-1.34	0.1861

c. **All regressions selections using Mallows CP statistic or adjusted-rsquare or RMSE. See below for comment and list of equations. Output is not necessary except the 3 summary models at the end of this section.**

Best 10 models by adjusted rsquare 09:39 Monday, April 29, 2002 377

The REG Procedure
Model: MODEL1
Dependent Variable: VAR4

Adjusted R-Square Selection Method

Number in Model	Adjusted R-Square	R-Square	C(p)	Variables in Model
5	0.3049	0.3704	2.9308	VAR3 VAR5 VAR7 NGEN VAR10
4	0.3035	0.3560	1.9565	VAR3 VAR5 NGEN VAR10
6	0.2989	0.3783	4.3731	VAR2 VAR3 VAR5 VAR7 NGEN VAR10
5	0.2981	0.3643	3.3679	VAR1 VAR3 VAR5 NGEN VAR10
6	0.2979	0.3773	4.4361	VAR1 VAR3 VAR5 VAR7 NGEN VAR10
5	0.2928	0.3595	3.7101	VAR3 VAR5 VAR9 NGEN VAR10
6	0.2922	0.3723	4.7966	VAR3 VAR5 VAR7 VAR9 NGEN VAR10
5	0.2916	0.3586	3.7736	VAR2 VAR3 VAR5 NGEN VAR10
5	0.2906	0.3575	3.8538	VAR3 VAR5 VAR8 NGEN VAR10
6	0.2905	0.3706	4.9072	VAR3 VAR5 VAR7 VAR8 NGEN VAR10

Best models by adjusted rsquare 09:39 Monday, April 29, 2002 395

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	VAR1	VAR2	VAR3	VAR5
1	MODEL1	PARMS	VAR4	1.02595	16.6488	.	.	-0.25138	-0.57346
2	MODEL1	PARMS	VAR4	1.02697	15.7074	.	.	-0.24037	-0.55249
3	MODEL1	PARMS	VAR4	1.03035	17.1599	.	0.42666	-0.26010	-0.60441
4	MODEL1	PARMS	VAR4	1.03093	16.5889	-.007509228	.	-0.25629	-0.52941
5	MODEL1	PARMS	VAR4	1.03111	17.4119	-.006874239	.	-0.26544	-0.55136
6	MODEL1	PARMS	VAR4	1.03481	15.6706	.	.	-0.24137	-0.57229
7	MODEL1	PARMS	VAR4	1.03526	16.5756	.	.	-0.25158	-0.58711
8	MODEL1	PARMS	VAR4	1.03553	15.8768	.	0.23573	-0.24367	-0.56706

9	MODEL1	PARMS	VAR4	1.03644	16.4777	.	.	-0.25509	-0.56206
10	MODEL1	PARMS	VAR4	1.03653	16.9957	.	.	-0.25819	-0.57754

Obs	VAR7	VAR8	VAR9	NGEN	VAR10	VAR4	_IN_	_P_	_EDF_	_RSQ_	_CP_
1	-0.70785	.	.	-3.82349	1.87274	-1	5	6	48	0.37045	2.93079
2	.	.	.	-3.47333	1.19375	-1	4	5	49	0.35605	1.95847
3	-0.66143	.	.	-3.80486	1.98917	-1	6	7	47	0.37826	4.37311
4	.	.	.	-3.46045	1.15948	-1	5	6	48	0.36432	3.36785
5	-0.67482	.	.	-3.79536	1.81884	-1	6	7	47	0.37735	4.43807
6	.	.	0.26247	-3.31521	1.11929	-1	5	6	48	0.35963	3.71008
7	-0.67333	.	0.19449	-3.68924	1.76445	-1	6	7	47	0.37232	4.79885
8	.	.	.	-3.42109	1.17669	-1	5	6	48	0.35864	3.77363
9	.	-0.19646	.	-3.61559	1.19768	-1	5	6	48	0.35751	3.85365
10	-0.68937	-0.09427	.	-3.63473	1.85690	-1	6	7	47	0.37078	4.90715

Best three models using adjusted-rsquared.

1. **Posterior error**=16.6468-0.25138*'BASE CURVE'-0.57346*'CLEAR ZONE'-0.70785*'INCISION' -3.82349*'GENDER 1 m 0 f'+1.87274*'INTER GENDER INCISION'; AdjRsq=.305
2. **Posterior error**=15.7074-0.24037*'BASE CURVE'-0.55249*'CLEAR ZONE' -3.47333*'GENDER 1 m 0 f'+1.19375*'INTER GENDER INCISION'; AdjRsq=.304
3. **Posterior error**=17.1599+0.42686*'BASE ERROR' -0.26010*'BASE CURVE'-0.60441*'CLEAR ZONE'-0.66143 *'INCISION' -3.80486*'GENDER 1 m 0 f'+1.98917*'INTER GENDER INCISION'; AdjRsq=.299

One could look at CP statistics, also as in the program, or best RMSE which is equivalent to best adjusted-Rsq.

d. Best model using forward regression. Output is not necessary, just model at bottom for forward or backward.

Best model according to forward selection 09:39 Monday, April 29, 2002 368

Forward Selection: Step 6

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	30.35646	5.05941	4.77	0.0007
Error	47	49.89648	1.06163		
Corrected Total	53	80.25294			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	17.15993	5.01736	12.41802	11.70	0.0013
VAR2	0.42686	0.55539	0.62710	0.59	0.4460
VAR3	-0.26010	0.11056	5.87310	5.53	0.0229
VAR5	-0.60441	0.18683	11.11039	10.47	0.0022
VAR7	-0.86143	0.70727	1.57486	1.48	0.2293
NGEN	-3.80466	1.46593	7.15190	6.74	0.0126
VAR10	1.98917	0.88192	5.40078	5.09	0.0288

Posterior error=17.15993+0.42686 * 'BASE ERROR' -0.26010 * 'BASE CURVE' -0.60441 * 'CLEAR ZONE' -0.86143 * 'INCISION' -3.80466 * 'GENDER 1 m 0 f' +1.98917 * 'INTER GENDER INCISION'.

Best model using backward selection

Best model according to backward selection 09:39 Monday, April 29, 2002 430

The REG Procedure

Model: MODEL1

Dependent Variable: VAR4 POST ERROR

Backward Elimination: Step 5

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	28.57375	7.14344	6.77	0.0002
Error	49	51.67919	1.05468		
Corrected Total	53	80.25294			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	15.70739	4.87451	10.95129	10.38	0.0023
VAR3	-0.24037	0.10913	5.11622	4.85	0.0324
VAR5	-0.55249	0.18074	9.85535	9.34	0.0036
NGEN	-3.47333	1.42211	6.29138	5.97	0.0182
VAR10	1.19375	0.57370	4.56643	4.33	0.0427

Posterior error=15.70739 -0.24037 * 'BASE CURVE' -0.55249 * 'CLEAR ZONE' -3.47333 * 'GENDER 1 m 0 f' +1.19375 * 'INTER GENDER INCISION'.

e. **Model similarities and differences. The full model is a very poor idea, especially if it includes subject number as above. There is no reason to subject identification as an independent variable, usually. However, the best models using selection methods are not identical, but have a lot in common. I have listed five examples below. The constant terms and variables included always consist of base curve, gender and interaction between gender and incision. Three of the five include incision and two also include base error.**

Forward **Posterior error**=17.15993+0.42686 * 'BASE ERROR' -0.26010 * 'BASE CURVE' -0.60441 * 'CLEAR ZONE' -0.86143 * 'INCISION' -3.80486 * 'GENDER 1 m 0 f' +1.98917 * 'INTER GENDER INCISION'

Backward: **Posterior error**=15.70739 -0.24037 * 'BASE CURVE' -0.55249 * 'CLEAR ZONE' -3.47333 * 'GENDER 1 m 0 f' +1.19375 * 'INTER GENDER INCISION'.

Best three models using adjusted-rsquared.

Posterior error=16.6488-0.25138*'BASE CURVE'-0.57346*'CLEAR ZONE'-0.70785*'INCISION' -3.82349*'GENDER 1 m 0 f'+1.87274*'INTER GENDER INCISION'; AdjRsq=.305

Posterior error=15.7074-0.24037*'BASE CURVE'-0.55249*'CLEAR ZONE' -3.47333*'GENDER 1 m 0 f'+1.19375*'INTER GENDER INCISION'; AdjRsq=.304

Posterior error=17.1599+0.42686*'BASE ERROR' -0.26010*'BASE CURVE'-0.60441*'CLEAR ZONE'-0.86143 *'INCISION' -3.80486*'GENDER 1 m 0 f'+1.98917*'INTER GENDER INCISION'; AdjRsq=.299

Solution #4 OB

a) $Y \sim Bin(n, p_0)$ $\qquad \frac{Y}{n}$

$\mu_Y = E[Y] = np_0$ $\qquad E[\frac{Y}{n}] = p_0$

$\sigma_Y^2 = V(Y) = np_0(1-p_0)$ $\qquad V(\frac{Y}{n}) = \frac{p_0(1-p_0)}{n}$

$$\tilde{\lambda} = -\log\left(\frac{Y}{n}\right)$$

$$\tilde{\lambda} = g(Y) = -\log\left(\frac{Y}{n}\right).$$

$$g'(Y) = -\frac{1}{Y/n}\cdot\frac{1}{n} = -\frac{1}{Y}$$

δ-method.
$Y = g(X)$
$\mu_Y \approx g(\mu_X)$
$\sigma_Y^2 \approx \sigma_X^2 [g'(\mu_X)]^2$

δ-method.

$$E[\tilde{\lambda}] \approx g(np_0) = -\log\left(\frac{np_0}{n}\right) = -\log(p_0)$$

$$V(\tilde{\lambda}) \approx np_0(1-p_0)\left[-\frac{1}{np_0}\right]^2 = \frac{(1-p_0)}{n\,p_0}$$

b) $$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$L(\lambda) = \prod_{i=1}^{n} f(x_i|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

$$= \left(\prod_{i=1}^{n} x_i!\right)^{-1} e^{-\lambda n}\lambda^{\Sigma x_i}$$

$$\ell(\lambda) = -\log\left(\prod_{i=1}^{n} x_i!\right) - \lambda n + \Sigma x_i \log(\lambda)$$

$$\ell'(\lambda) = -n + \frac{\Sigma x_i}{\lambda} = 0$$

$$\frac{\Sigma x_i}{\lambda} = n.$$

$$\boxed{\hat{\lambda} = \frac{\Sigma x_i}{n} = \bar{x}}$$

c) $$Var(\hat{\lambda}) = Var(\bar{X}) = Var\left(\frac{\Sigma x_i}{n}\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\lambda = \frac{\lambda}{n}.$$

d) $$eff(\tilde{\lambda}, \hat{\lambda}) = \frac{Var(\hat{\lambda})}{Var(\tilde{\lambda})} = \frac{\frac{\lambda}{\cancel{n}}}{\frac{(1-p_0)}{\cancel{n}\, p_0}} = \frac{\lambda p_0}{(1-p_0)}$$

$$= \frac{\lambda}{\frac{1}{p_0} - 1} = \frac{\lambda}{e^{\lambda} - 1} < 1$$

since $$e^{\lambda} = \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} > 1 + \lambda.$$

$\therefore$ MLE is more efficient.

e) $$\hat{\lambda} \pm z_{\alpha/2} \frac{1}{\sqrt{n I(\hat{\lambda})}}$$

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

$$= -E\left[-\frac{x}{\lambda^2}\right].$$

$$= \frac{1}{\lambda^2}\cdot\lambda = \frac{1}{\lambda}$$

$$f(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

$$\log f(x|\lambda) = -\lambda + x\log(\lambda) - \log(x!)$$

$$\frac{d}{d\lambda}\log f(x|\lambda) = -1 + \frac{x}{\lambda}$$

$$\frac{d^2}{d\lambda^2}\log f(x|\lambda) = -\frac{x}{\lambda^2}$$

large sample $100(1-\alpha)\%$ CI for λ

$$\hat{\lambda} \pm z_{\alpha/2} \frac{1}{\sqrt{n \cdot \frac{1}{\hat{\lambda}}}}$$

$$\boxed{\hat{\lambda} \pm z_{\alpha/2} \sqrt{\frac{\hat{\lambda}}{n}}}$$

f) i) Using the estimator $\tilde{\lambda}$:

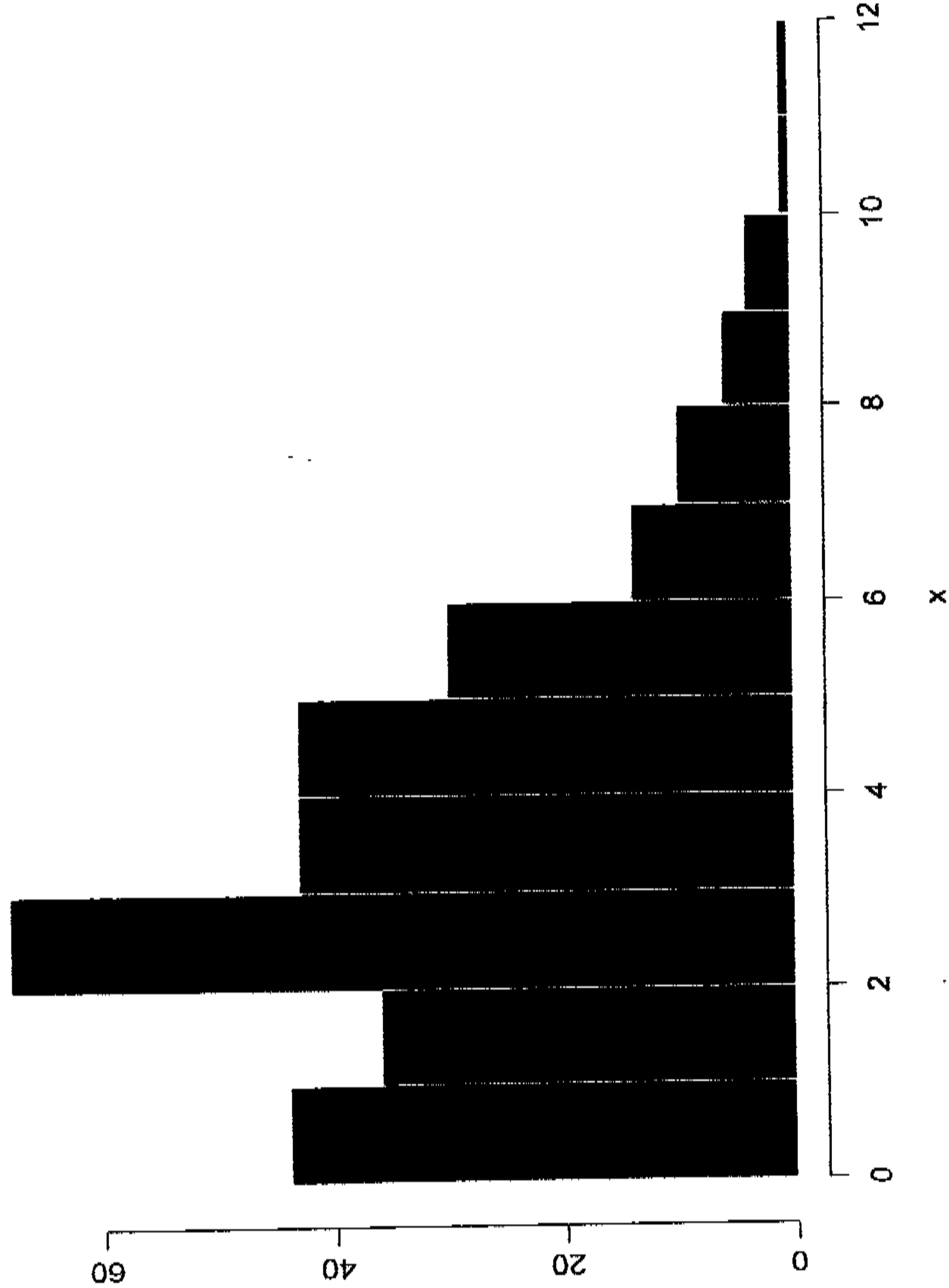
1. compute $\tilde{\lambda}$ for the original data $x_1, \ldots, x_n$.
2. using $\tilde{\lambda}$, sample B bootstrap samples of size n from $Poi(\tilde{\lambda})$. $\underset{\sim}{X}_1^*, \underset{\sim}{X}_2^*, \ldots, \underset{\sim}{X}_B^*$.
3. Compute $\tilde{\lambda}^*$ for each bootstrap sample.
4. using the percentile method

$$(\tilde{\lambda}^*_{(.025)}, \tilde{\lambda}^*_{(.975)})$$

gives a 95% bootstrap CI for λ

ii) Using the MLE $\hat{\lambda}$:

same as above substituting $\hat{\lambda}$ for $\tilde{\lambda}$.


0
2
4
6
8
10
12
x
0
20
40
60