STATISTICS DEPARTMENT
M.S. EXAMINATION

PART II
OPEN BOOK

Tuesday, May 20, 2003

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 weights of a random sample of 24 male runners are measured. The sample mean is 60 kilograms (kg). Suppose that the standard deviation is known to be 5 kg.

(a) Describe an appropriate population and sample for this problem. That is, tell the story of this experiment so that it can be analyzed correctly. Name the statistics and parameters mentioned and say whether these are known or unknown. [2 points]

(b) What is the standard error for 60 kg? [2 points]

(c) Is it appropriate to use the Central Limit Theorem here? Why or why not? If it is appropriate, how does the CLT apply? To what does it apply? [4 points]

(d) Give a 95% confidence interval for the mean of the population from which the sample is drawn. [4 points]

(e) Because Americans are less familiar with kilograms, convert to pounds by multiplying by the conversion factor 2.2 pounds per kilogram. What are the new values for the sample mean and standard error of the mean when measured in pounds? [3 points]

(f) In a new sample and with weights measured in pounds, what sample size n would we need to estimate the population mean using a 95%-confidence interval which is centered within a margin of error of 1.5 pounds? [5 points]

(g) Design a test of the hypothesis that the population mean for male runners is under 130 pounds. Use a Type I error of 0.05 and select n so that, if the true mean is 128 pounds, the Type II error is approximately 0.05 as well. [5 points]

2. Two types of laboratory tests (A and B) are used to determine the level of a certain liver enzyme in human blood. It is claimed that both tests accurately measure this enzyme. The question is whether they actually give equivalent results. Suppose that a study to investigate this is conducted at three randomly chosen hospitals. At each hospital blood is drawn from 20 randomly chosen subjects (60 subjects in all). The blood from each subject is divided into two samples, one assayed for the enzyme using each type of test (120 assays in all).

Suppose the enzyme levels obtained are as shown in the table below. For each type of test, the results from the 20 subjects at a hospital are given in the same order across two rows of 10 numbers each. These data are available in the order shown below (but without labels) in the file ENZYME.TXT. They are also displayed in two-column format in ENZYME2.TXT and in a Minitab worksheet ENZYME2.MTW.

TYPE A

Hospital										
1	154	165	149	144	139	160	154	150	146	146
	154	155	143	150	166	157	165	136	149	139
2	152	158	151	157	146	143	143	149	162	152
	136	154	149	150	136	159	155	137	159	157
3	145	135	163	148	152	158	141	159	138	144
	145	158	133	150	147	157	152	152	137	152

TYPE B

Hospital										
1	150	141	160	147	125	144	158	142	135	152
	155	143	165	149	156	153	127	153	146	168
2	128	146	133	135	132	146	128	142	139	156
	132	124	127	128	140	149	133	133	134	129
3	149	152	158	152	167	162	156	159	156	141
	139	165	150	146	155	159	149	157	142	138

(a) Write the most complete ANOVA model supported by these data. Account for all possible interactions. Say whether each main effect is fixed or random, and how many levels there are. If there is nesting, describe it. Also state the assumptions of your model.

(b) Perform the numerical analysis according to your model. Give a table with columns headed Source, DF, SS, MS, F, and P. Discuss any significant effects, explaining their meaning in nontechnical language that could be understood by someone with no background in ANOVA designs.

(c) Perform the appropriate procedures to check assumptions and report your findings.

(d) Would it make any difference in the F-ratios if you changed the model designation of the hospital effect—fixed vs. random? Would this change make any difference in the practical interpretation of your results? Explain.

(e) These are *fake* data. They fail to exhibit an important property that one would expect to see very clearly in *real* data collected according to the story above. Identify what is missing and discuss. (If this were a living room, we would be talking about an elephant lounging on the sofa, not some dust on the coffee table.)

3. Consider a data set containing the cumulative GPA for a random sample of computer science majors at a large university. This data set is located in the text-file GRADES. There are several explanatory variables including High School Mathematics grade (1-10), High School Social Science grade (1-10), High School English grade (1-10), SAT mathematics (1-800), and SAT verbal scores (1-800). Gender is also recorded (m or f). The first few lines of data are as follows:

001	3.32	10	10	10	670	600	m
002	2.26	6	8	5	700	640	m
003	2.35	8	6	8	640	530	m
004	2.08	9	10	7	670	600	m
005	3.38	8	9	8	540	580	m
006	3.29	10	8	8	760	630	m
007	3.21	8	8	7	600	400	m
008	2.00	3	7	6	460	530	m
009	3.18	9	10	8	670	450	m
010	2.34	7	7	6	570	480	m

(a) Read in the file GRADES using a SAS program. (4 pts.)

(b) Ignoring gender, create a model for predicting college GPA containing all 5 other explanatory variables. (4 pts.)

(c) Again ignoring gender, create a smaller model for college GPA containing a subset of the 5 explanatory variables. Describe the method you used to choose this model. Is it is better or worse than the model in (b). (5 pts.)

(d) Discuss the model assumptions using the residuals from (c). Include statistics, hypothesis test(s), and at least one graph that is relevant to model assessment. (6 pts.)

(e) Include gender in the model. Indicate whether the model is improved and whether it is sensible to include an interaction with gender. Discuss why you think this might be true. (6pts.)

4. Consider the number of eggs the Queen Bee lays in a bee hive. Suppose the distribution of the random variable Y = the number of eggs laid by the Queen Bee is $Poisson(\lambda)$. Also suppose the random variable X = number of survivors is of interest to a Biologist.

   A hierarchical model for the number of survivors in terms of the number of eggs laid is defined as:

$$X|Y \sim Binomial(Y, p) \tag{1}$$

   and

$$Y \sim Poisson(\lambda) \tag{2}$$

   where p is the proportion of eggs that result in a surviving bee.

   (a) What is the expected value of Y? What is the variance of Y?

   (b) Early in the spring the Queen Bee may lay several hundred eggs per day. Suppose $\lambda = 300$. Compute the probability that the number of eggs laid is greater than 320, $P(Y > 320)$. Because λ is so large, what distribution could be used to approximate this probability calculation? Using that distribution compute the approximate probability of $P(Y > 320)$.

   (c) What is the expected value of $X|Y = y$? What is the variance of $X|Y = y$?

   (d) Assuming the survival rate of bees is $p = 0.9$ and the number of eggs laid is $Y = 300$, compute the probability that the number of survivors is greater than 280, $P(X > 280|Y = 300)$. Since Y is so large, what distribution could be used to approximate this probability calculation?

   (e) Suggest a different hierarchical model for the number of survivors in terms of the number of eggs laid considering the large sample distributions.

5) Suppose Factor A is fixed with 2 levels, Factor B (nested in A) is random with 3 levels and 3 observations are taken at each of the 6 combinations of A and B.

This model is usually written as $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$ $(i = 1, 2; j, k = 1, 2, 3)$.

(a) What are the usual assumptions for this model?

(b) Let $Y_{1\bullet\bullet} = \sum_{j=1}^{3}\sum_{k=1}^{3} Y_{1jk}$. Show that $Y_{1\bullet\bullet} = 9\mu + 9\alpha_1 + 3(\beta_{1(1)} + \beta_{2(1)} + \beta_{3(1)}) + \varepsilon_{1\bullet\bullet}$, where $\varepsilon_{1\bullet\bullet} = \sum_{j=1}^{3}\sum_{k=1}^{3} \varepsilon_{1jk}$.

(c) Show that $\text{var}(Y_{1\bullet\bullet}) = 9(3\sigma_\beta^2) + 9\sigma^2$, where σ_β^2 is the common variance of $\{\beta_{j(i)}\}$ and σ^2 is the common variance of $\{\varepsilon_{ijk}\}$.

(d) Letting $\bar{Y}_{1\bullet\bullet} = \frac{Y_{1\bullet\bullet}}{9}$, show that $\text{var}(\bar{Y}_{1\bullet\bullet}) = \frac{1}{9}[3\sigma_\beta^2 + \sigma^2]$.

(e) Obtain $E(\bar{Y}_{1\bullet\bullet} - \bar{Y}_{2\bullet\bullet})$ and $\text{var}(\bar{Y}_{1\bullet\bullet} - \bar{Y}_{2\bullet\bullet})$.

(f) Suppose we compare 2 drugs, with 3 randomly selected batches from each drug. We randomly select 3 individuals for each combination of drug and batch; measure Y = improvement for each individual. The data are given below. **The corresponding file is named "improvement" and it is saved as Minitab, Excel and 'dat' files.** For A = drug and B = batch, from the ANOVA table we obtain $E(\text{MS[B(A)]}) = 3\sigma_\beta^2 + \sigma^2$, df[B(A)] = 4, MS[B(A)] = .366. Test H_0: $\alpha_1 = \alpha_2$.

(g) Explain in words, without any technical jargon, what conclusion can be drawn based on the result of the hypothesis test in part (f).

obs.	drug	batch	improvement
1	1	1	1.257
2	1	1	1.415
3	1	1	2.172
4	1	2	2.743
5	1	2	2.250
6	1	2	2.179
7	1	3	1.000
8	1	3	1.657
9	1	3	2.107
10	2	1	6.007
11	2	1	6.457
12	2	1	5.329
13	2	2	5.936
14	2	2	6.493
15	2	2	5.693
16	2	3	6.857
17	2	3	5.550
18	2	3	6.500

Part II

Solution #1

1. The weights of a random sample of 24 male runners are measured. The sample mean is 60 kilograms (kg). Suppose that the standard deviation is known to be 5 kg.

a. Describe an appropriate population and sample for this problem. Name the statistics and parameters mentioned and whether these are known or unknown. (2pts.)

The population is all male runners completing marathons in top fifty, worldwide in 2002. The sample is a randomly selected sample taken from records of all marathon runners who were male. The statistic given is that the sample weight, $\bar{x}$ **is 60 kg. The standard deviation 5 kg is the population parameter σ.**

b. What is the standard error for 60 kilograms? (2pts.)

$\sigma_{\bar{x}} = 5/\sqrt{24}$ **= 1.02**

c. Why or why not is it appropriate to use the Central Limit Theorem here? How and to what does the CLT apply. (3pts.)

We are given σ, therefore μ exists. We have a sample of size 24 that is fairly large. We don't know the distribution of the original measurements. But we can assume that the sample mean is approximately normal using the CLT, so we assume that $\bar{x}$ is approximately normal with unknown mean μ and standard deviation 1.02.

d. Give a 95% confidence interval for the mean of the population from which the sample is drawn. (3pts.)

Applying the CLT, a 95% confidence interval for μ is the interval 60±1.96*1.02 in kilograms.

e. Since Americans are less familiar with kilograms, we wish to convert to pounds by multiplying by the conversion factor 2.2 pounds per kilogram. What are the new values for mean and standard error of the mean measurements in pounds? (3pts.)

$\bar{x}$***2.2=132 pounds with standard error 2.2*** $\sigma_{\bar{x}} = 2.2*5/\sqrt{24}$ **= 2.2*1.02=2.24 pounds**

f. What n would we need in a new sample to estimate the mean using a 95% confidence interval which when measured in pounds is centered within 1.5 pounds of the mean? (3pts.)

1.5=1.96*2.2*5/√n so n must be 207.

g. Design a hypothesis test for the mean, measured in pounds, that asserts the population mean for male runners is under 130 pounds. Use a type I error of .05 and select n so that if the true mean is 128 pounds, the type II error is approximately .05 as well. (4pts.)

The critical value for a .05 test for μ would be 130-1.645*2.2*5/√n. Values above this one would indicate fail to reject Ho and values below would indicate reject Ho. If the true μ is 128, then we would tend to observe values above 130-1.645*2.2*5/√n with probability set at .05. Converting back to a standard normal we have (2-1.645*2.2*5/√n)/ 2.2*5/√n =1.645. Solve for n. n=328.

Answers

(a) Model: $Y_{ijk} = \mu + H_i + S(H)_{j(i)} + \tau_k + (H\tau)_{ik} + e_{ijk}$, where $i = 1,2$, $j = 1, ...,20$, $k = 1,2,3$. Distributions: H_i iid $N(0, \sigma_H^2)$, $S(H)_{j(i)}$ iid $N(0, \sigma_S^2)$, e_{ijk} iid $N(0, \sigma^2)$. Optionally, we use the restricted model here, so that $\Sigma_k (H\tau)_{ik} = 0$, but similar results are obtained without this assumption. Main effects: Hospitals random, Subjects random and nested within Hospitals, and test Types fixed. (Major points off for omitting interaction; fixed/random, crossed/nested errors. Minor points off for skipping other details.)

(b)

Analysis of Variance for Enz

Source	DF	SS	MS	F	P
Type	1	550.41	550.41	0.61	0.518
Hosp	2	1363.55	681.78	7.83	0.001
Type*Hosp	2	1819.12	909.56	12.46	0.000
Subj(Hosp)	57	4960.88	87.03	1.19	0.254
Error	57	4159.98	72.98		
Total	119	12853.93			

Source	Variance component	Error term	Expected Mean Square for Each Term (using restricted model)
1 Type		3	(5) + 20(3) + 60Q[1]
2 Hosp	14.869	4	(5) + 2(4) + 40(2)
3 Type*Hosp	41.829	5	(5) + 20(3)
4 Subj(Hosp)	7.025	5	(5) + 2(4)
5 Error	72.982		(5)

Hospital*Type very highly significant. Also disorderly. We can take no comfort in failure to reject the Type effect, test methods may be implemented differently at different hospitals. (Major points off for incorrect interpretation of main effects ignoring interaction.)

Rows: Hosp Columns: Type

	1	2	All
1	151.05	148.45	149.75
2	150.25	135.70	142.98
3	148.30	152.60	150.45
All	149.87	145.58	147.73

Cell Contents --
Enz:Mean

(c) A normal probability plot of residuals from the model seems satisfactorily close to linear. Also, a plot of residuals in the order shown in the problem reveals no heteroscedastic pattern. One could also do formal tests. (Major points off for checking neither normality nor homoscedasticity; minor penalty for skipping one.)

(d) Yes, the F ratios are different. The interaction and the small number of hospitals keeps Type from having a small P-value (tested against interaction). But if we take the hospitals to be fixed effects, then Type has a very small P-value (tested against error). [Unrestricted model: Random Hospital requires synthetic test, Type not significant; fixed hospital gives same F-ratios as for restricted model.]

Analysis of Variance for Enz

Source	DF	SS	MS	F	P
Type	1	550.41	550.41	7.54	0.008
Hosp	2	1363.55	681.78	7.83	0.001
Type*Hosp	2	1819.12	909.56	12.46	0.000
Subj(Hosp)	57	4960.88	87.03	1.19	0.254
Error	57	4159.98	72.98		
Total	119	12853.93			

Source	Variance component	Error term	Expected Mean Square for Each Term (using restricted model)
1 Type		5	(5) + 60Q[1]
2 Hosp		4	(5) + 2(4) + 40Q[2]
3 Type*Hosp		5	(5) + 20Q[3]
4 Subj(Hosp)	7.025	5	(5) + 2(4)
5 Error	72.982		(5)

However, the disorderly interaction appears in all cases, preventing a clear interpretation of either main effect, so the real-world interpretation is somewhat similar whether hospitals are taken as fixed or random [also restricted or unrestricted].

(e) In *any* experiment with randomly chosen human subjects, one expects the Subject effect to be very highly significant. Otherwise, why use so many people?! Here, it isn't anywhere near significant. (Elephants are large so they can look different from different angles. Equivalently, observe that, for each of the three hospitals, the 20 paired A and B measurements have no significant correlation.) [A related issue: anyone who knows about liver enzymes would be astonished not to find among the 60 randomly chosen subjects a few outliers with A and B assays both very high. Result: both correlation and nonnormal residuals.]

Solution #3

4. Consider a data set containing the cumulative GPA for a random sample of computer science majors at a large university. This data is located in the file GRADES. There are several explanatory variables including High School Mathematics grade, High School Social Science grade, and High School English grade, SAT Mathematics and SAT verbal scores. Gender is also recorded.

a. Read in the file Grades using a SAS program. (3 pts.)

FOR EXAMPLE:

```
data grades;
infile 'c:/temp/CSDATA for regression.txt';
input student gpa HSMATH HSSS HSENG SATMATH SATVERB gender $;
title 'problem a';
/*
001    3.32     10     10     10     670     600     m
002    2.26      6      8      5     700     640     m
003    2.35      8      6      8     640     530     m
004    2.08      9     10      7     670     600     m
005    3.38      8      9      8     540     580     m
006    3.29     10      8      8     760     630     m

*/;
```

b. Ignoring gender, create a model for college GPA containing all 5 explanatory variables. (3 pts.)

Using the following code we get the solution below from SAS:

```
title 'problem b';
proc reg;
model gpa=HSMATH HSSS HSENG SATMATH SATVERB/press r; run;
```

problem b 18:04 Monday, May 19, 2003 1

The REG Procedure
Model: MODEL1
Dependent Variable: gpa

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	28.64364	5.72873	11.69	<.0001
Error	218	106.81914	0.49000		
Corrected Total	223	135.46279			

Root MSE	0.70000	R-Square	0.2115
Dependent Mean	2.63522	Adj R-Sq	0.1934

Coeff Var 26.56311

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.32672	0.40000	0.82	0.4149
HSMATH	1	0.14596	0.03926	3.72	0.0003
HSSS	1	0.03591	0.03780	0.95	0.3432
HSENG	1	0.05529	0.03957	1.40	0.1637
SATMATH	1	0.00094359	0.00068566	1.38	0.1702
SATVERB	1	-0.00040785	0.00059189	-0.69	0.4915

Therefore the model is:
gpa=.327+.146*HSMATH+.036*HSSS+.055*HSENG+,00094*SASMATH-.00041*SATVERB

c. Again ignoring gender, create a smaller model for college GPA containing a subset of the 5 exploratory variables. Describe the method you used to choose this model and whether it is better or worse than the model in b. (4 pts.)

Using the code below or some other similar selection method:

```
proc reg;
model gpa=HSMATH HSSS HSENG SATMATH SATVERB/selection=cp rmse adjrsq;
title 'problem c'; run;
proc reg;
model gpa=HSMATH HSENG SATMATH/press r;
output out=resids student=student p=fits;
proc univariate data=resids plot normal;
var student; run;
```

According to the output below, I would choose model 2 and include HSMATH, HSENG, and SATMATH in the model. Although none of these models are very good, this one has the best adjusted r-squared. Model 1 has all coefficients significant while this 3 variable model does not. Only 11 of the 172 studentized residuals are outside ± 2 and none are outside ± 3. Three of the four normal tests reject normal errors, however. Even so, this is a simpler model than the one in part b. The smaller model also has a lower press statistic indicating better prediction.

oblem c 18:04 Monday, May 19, 2003 2

The REG Procedure
Model: MODEL1
Dependent Variable: gpa

C(p) Selection Method

Number in Model	C(p)	R-Square	Adjusted R-Square	Root MSE	Variables in Model
2	2.7350	0.2016	0.1943	0.69958	HSMATH HSENG
3	3.2512	0.2069	0.1961	0.69880	HSMATH HSENG SATMATH
2	3.2585	0.1997	0.1924	0.70041	HSMATH HSSS
1	3.7832	0.1905	0.1869	0.70280	HSMATH
3	3.9007	0.2046	0.1937	0.69984	HSMATH HSSS HSENG
3	4.1598	0.2036	0.1928	0.70025	HSMATH HSSS SATMATH
4	4.4748	0.2097	0.1953	0.69916	HSMATH HSSS HSENG SATMATH
3	4.7348	0.2016	0.1907	0.70117	HSMATH HSENG SATVERB
2	4.7775	0.1942	0.1869	0.70281	HSMATH SATMATH
4	4.9023	0.2082	0.1937	0.69984	HSMATH HSENG SATMATH SATVERB
3	5.2570	0.1997	0.1888	0.70199	HSMATH HSSS SATVERB
2	5.6893	0.1909	0.1835	0.70424	HSMATH SATVERB
4	5.8939	0.2046	0.1901	0.70142	HSMATH HSSS HSENG SATVERB
4	5.9527	0.2044	0.1899	0.70152	HSMATH HSSS SATMATH SATVERB
5	6.0000	0.2115	0.1934	0.70000	HSMATH HSSS HSENG SATMATH SATVERB
3	6.7619	0.1942	0.1832	0.70438	HSMATH SATMATH SATVERB
3	17.2321	0.1564	0.1448	0.72074	HSSS HSENG SATMATH
4	17.8214	0.1615	0.1461	0.72020	HSSS HSENG SATMATH SATVERB
2	19.7248	0.1401	0.1323	0.72600	HSSS SATMATH
3	20.9845	0.1428	0.1311	0.72652	HSSS SATMATH SATVERB
2	21.7757	0.1327	0.1248	0.72913	HSENG SATMATH
3	22.7150	0.1365	0.1247	0.72916	HSENG SATMATH SATVERB
2	24.4473	0.1230	0.1151	0.73318	HSSS HSENG
3	26.3825	0.1233	0.1113	0.73474	HSSS HSENG SATVERB
1	26.4555	0.1085	0.1045	0.73755	HSSS
2	28.2181	0.1094	0.1013	0.73886	HSSS SATVERB
1	33.3667	0.0835	0.0794	0.74782	HSENG
2	34.7962	0.0856	0.0773	0.74866	HSENG SATVERB
1	38.9406	0.0634	0.0591	0.75600	SATMATH
2	40.9387	0.0634	0.0549	0.75770	SATMATH SATVERB
1	52.8331	0.0131	0.0087	0.77601	SATVERB

d. Discuss the model assumptions using the residuals from c. Include statistics, hypothesis test(s), and at least one graph that is relevant to model assessment. (5 pts.)

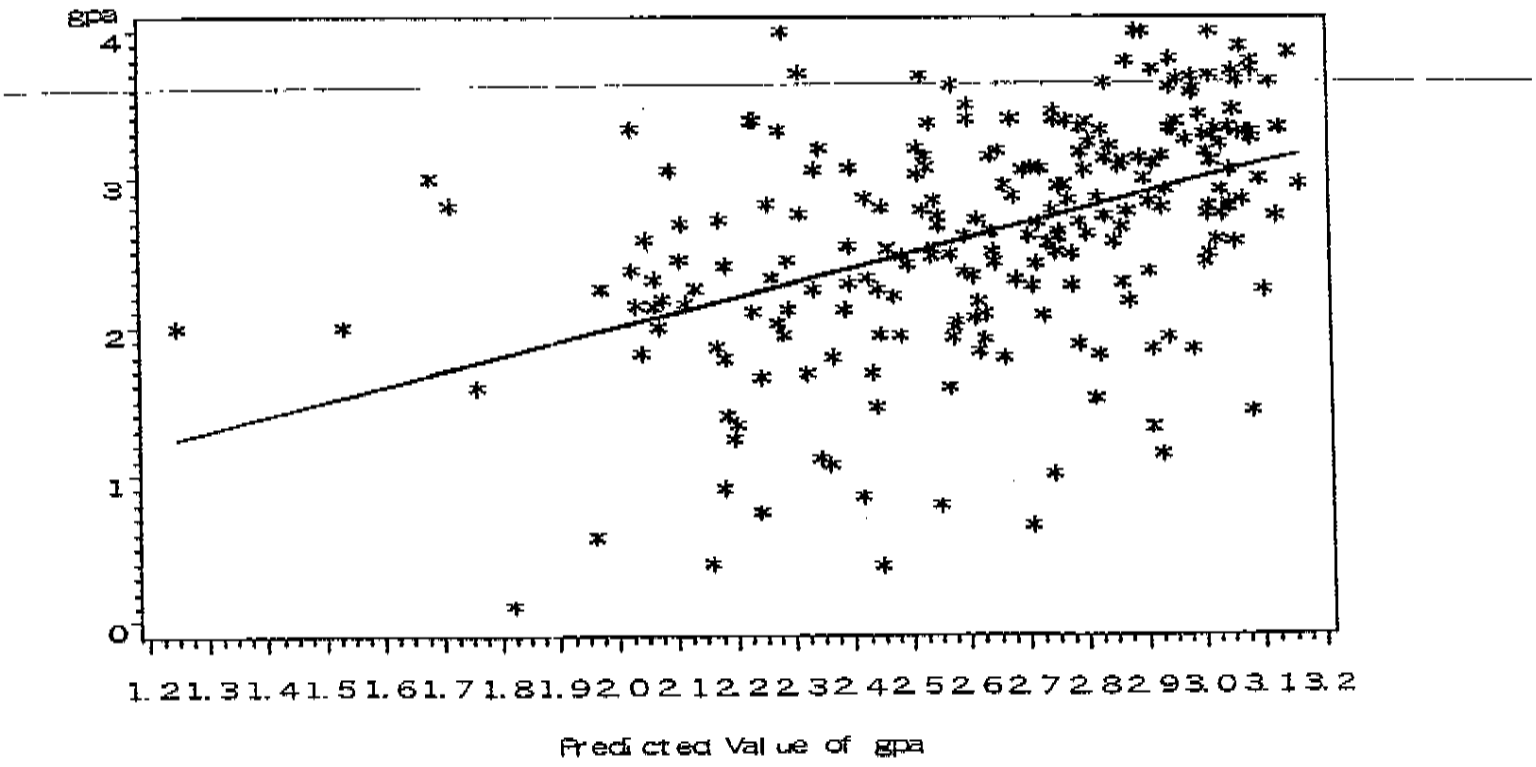
A number of ideas could be used here such as the 4 normal tests, the press statistic for comparison, etc.

Several useful graphs and tests come from the code below:

```
proc univariate data=resids plot normal;
var student;run;

proc gplot data=resids;
symbol1 v=star cv=blue i=none;
symbol2 v=none i=line ci=black;
plot gpa*fits=1 fits*fits=2/overlay;
title 'problem d';
```

Solid line shows fit versus fit. A model with nearly perfect fit would fall much nearer the line. Some important information is missing from this model.



e. Include gender in the model. Indicate whether the model is improved and whether it is sensible to include an interaction with gender and why you think this might be true. (5pts.)

Several models fit approximately the same. Again, I will choose the model with HSENG HSMATH SATMATH plus sex. We have the same problem with lack of normality. We have lowered the Press error so improved predictability somewhat. Many reasonable comparisons could be made--but we are still missing a big part of the picture. Adjusted r-square has increased slightly to .20 and three of the four coefficients are significant. SATMATH is not significant with these other terms in the equation.

One could also consider heteroscedasticity or multicolinearity or including powers or interaction terms. The best interaction term is sex*satmath which increases the adjusted r-squared to .2053 with all three coefficients significant. But there are still problems with the residuals being non-normal by three of the four tests.

When squares of the hs grades are included, hsss becomes important:

C(p) Selection Method

Number in Model	C(p)	R-Square	Adjusted R-Square	Root MSE	Variables in Model
4	1.6853	0.2505	0.2368	0.68089	HSSS sexhseng hsmathsq hssssq
3	1.8439	0.2430	0.2327	0.68273	HSSS hsmathsq hssssq
4	1.9032	0.2497	0.2360	0.68123	HSSS sex hsmathsq hssssq
4	2.0149	0.2494	0.2356	0.68141	HSSS sexsatmath hsmathsq hssssq
5	2.0661	0.2561	0.2391	0.67988	HSSS sex sexhsmath hsmathsq hssssq
5	2.0977	0.2560	0.2389	0.67993	HSSS HSENG sex hsmathsq hssssq
4	2.1007	0.2491	0.2353	0.68154	HSSS sexhsss hsmathsq hssssq
6	2.2298	0.2625	0.2421	0.67852	HSSS HSENG sex sexhsmath hsmathsq hssssq
5	2.2732	0.2554	0.2383	0.68021	HSSS HSENG sexsatmath hsmathsq hssssq

Using the first model here with HSSS and hssssq as well as sexhseng and hsmathsq, the adjusted r-squared is raised to .237. The other difficulties are not removed such as lack of normality. The press statistic is reduced to 106 indicating even more predictability. A possible program follows:

```
data grades;
infile 'c:/temp/CSDATA for regression.txt';
input student gpa HSMATH HSSS HSENG SATMATH SATVERB gender $;
title 'problem a';
/*
001     3.32      10      10      10      670      600      m
002     2.26       6       8       5      700      640      m
003     2.35       8       6       8      640      530      m
004     2.00       5      10       7      670      600      f

223     2.59       5       4       7      630      470      f
224     2.28       3       3       5      550      460      f
*/;
title 'problem b';
proc reg;
model gpa=HSMATH HSSS HSENG SATMATH SATVERB/press r; run;
```

```
    reg;
 el gpa=HSMATH HSSS HSENG SATMATH SATVERB/selection=cp rmse adjrsq;
proc reg;
odel gpa=HSMATH HSENG SATMATH/press r;
utput out=resids student=student p=fits;
itle 'problem c';

roc univariate data=resids plot normal;
ar student;run;
roc gplot data=resids;
ymbol1 v=star cv=blue i=none;
ymbol2 v=none i=line ci=black;
lot gpa*fits=1 fits*fits=2/overlay;
itle 'problem d'; run;

data grades; set grades; sex=0;  if gender='m' then sex=1;
sexhsmath=sex*hsmath; sexhseng=sex*hseng; sexhsss=sex*hsss;
sexsatmath=sex*satmath; sexsatverb=sex*satverb; satmathverb=satmath*satverb;
smathsq=hsmath*hsmath; hsengsq=hseng*hseng; hssssq=hsss*hsss;
run;
proc reg;
model gpa=HSMATH HSSS HSENG SATMATH SATVERB sex/selection=cp rmse adjrsq; run;
proc reg;
model gpa=HSMATH HSENG SATMATH sex/press r;
output out=resids student=student p=fits;
title 'problem e';

p  univariate data=resids plot normal;
   student;run;
  c gplot data=resids;
plot gpa*fits=1 fits*fits=2/overlay; run;

proc reg;
model gpa=HSMATH HSSS HSENG SATMATH SATVERB sex
sexhsmath sexhseng sexhsss sexsatmath sexsatverb/selection=cp rmse adjrsq; run;
proc reg;
model gpa=HSMATH HSENG sexSATMATH/press r;
output out=resids student=student p=fits;
title 'problem e including interaction';

proc univariate data=resids plot normal;
var student;run;
proc gplot data=resids;
plot gpa*fits=1 fits*fits=2/overlay; run;

proc reg;
model gpa=HSMATH HSSS HSENG SATMATH SATVERB sex
sexhsmath sexhseng sexhsss sexsatmath sexsatverb
satmathverb hsmathsq hsengsq hssssq/selection=cp rmse adjrsq; run;
title 'problem e including interaction and squares';
proc reg;
model gpa=HSSS sexhseng hsmathsq hssssq/press r;
output out=resids student=student p=fits;
t  le 'problem e including interaction and squares';

  c univariate data=resids plot normal;
 ar student;run;
proc gplot data=resids;
plot gpa*fits=1 fits*fits=2/overlay;  run;
```

Solution # 4

4.

a. The mean and variance of the Poisson distribution is lambda.

```
mu.y <- lambda
var.y <- lambda
```

b. Using Splus.

```
# exact calculation

lambda <- 300

1 - ppois(320, lambda = 300)
```

Ans: 0.1190045

```
# normal approximation

mu.y <- lambda
var.y <- lambda
sigma.y <- sqrt(lambda)

1 - pnorm(320, mu.y, sigma.y)
```

Ans: 0.1241065

c.

```
mu.xgy <- y*p
var.xgy <- y*p*(1-p)
```

d.

```
# exact calculation

p <- 0.9
y <- 300

1 - pbinom(280,y,p)
```

Ans: 0.01711813

```
# normal approximation

mu.xgy <- y*p
var.xgy <- y*p*(1-p)
sigma.xgy <- sqrt(y*p*(1-p))

1 - pnorm(280,mu.xgy,sigma.xgy)
```

Ans: 0.02714591

e.

$Y \sim N(\mu_y = \lambda, \sigma^2_y = \lambda)$

$X|Y \sim N(\mu_{x|y} = yp, \sigma^2_{x|y} = yp(1-p))$

Solution #5

(5g) Suppose Factor A is fixed with 2 levels, Factor B (nested in A) is random with 3 levels and 3 observations are taken at each of the 6 combinations of A and B.

This model is usually written as $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$ $(i = 1, 2; j, k = 1, 2, 3)$.

(a) What are the usual assumptions for this model?

(b) Let $Y_{1..} = \sum_{j=1}^{3}\sum_{k=1}^{3} Y_{1jk}$. Show that $Y_{1..} = 9\mu + 9\alpha_1 + 3(\beta_{1(1)} + \beta_{2(1)} + \beta_{3(1)}) + \varepsilon_{1..}$, where $\varepsilon_{1..} = \sum_{j=1}^{3}\sum_{k=1}^{3} \varepsilon_{1jk}$.

(c) Show that $\text{var}(Y_{1..}) = 9(3\sigma_\beta^2) + 9\sigma^2$, where σ_β^2 is the common variance of $\{\beta_{j(i)}\}$ and σ^2 is the common variance of $\{\varepsilon_{ijk}\}$.

(d) Letting $\bar{Y}_{1..} = \frac{Y_{1..}}{9}$, show that $\text{var}(\bar{Y}_{1..}) = \frac{1}{9}[3\sigma_\beta^2 + \sigma^2]$.

(e) Obtain $E(\bar{Y}_{1..} - \bar{Y}_{2..})$ and $\text{var}(\bar{Y}_{1..} - \bar{Y}_{2..})$.

(f) Suppose we compare 2 drugs, with 3 randomly selected batches from each drug. We randomly select 3 individuals for each combination of drug and batch and measure Y = improvement for each individual. The data is given below. For A = drug, B = batch, from the ANOVA table, we obtain that $E(MS(B(A))) = 3\sigma_\beta^2 + \sigma^2$, df(B(A)) = 4, MS(B(A)) = .366. Test H_0: $\alpha_1 = \alpha_2$.

(g) In words, without any technical jargon, what conclusion can be made based on the result of the hypothesis test in part (f).

obs.	drug	batch	improvement
1	1	1	1.257
2	1	1	1.415
3	1	1	2.172
4	1	2	2.743
5	1	2	2.250
6	1	2	2.179
7	1	3	1.000
8	1	3	1.657
9	1	3	2.107
10	2	1	6.007
11	2	1	6.457
12	2	1	5.329
13	2	2	5.936
14	2	2	6.493
15	2	2	5.693
16	2	3	6.857
17	2	3	5.550
18	2	3	6.500

Solution (a) $\sum \alpha_i = 0$; $\{\beta_{j(i)}\}$ are normal with mean=0, common variance σ_β^2, mutually independent and independent of $\{\varepsilon_{ijk}\}$; $\{\varepsilon_{ijk}\}$ are normal with mean=0, common variance σ^2 and are mutually independent.

(b) by definition of Y_{ijk}

(c) from (b), assumptions and elementary properties of variance.

(d) from (c) and elementary properties of variance.

(e) $E(\bar{Y}_{1..} - \bar{Y}_{2..}) = (\mu + \alpha_1) - (\mu + \alpha_2) = \alpha_1 - \alpha_2$

$\text{var}(\bar{Y}_{1..} - \bar{Y}_{2..}) = \text{variance}(\bar{Y}_{1..}) + \text{var}(\bar{Y}_{2..})$

$= \frac{2}{9}[3\sigma_\beta^2 + \sigma^2]$

(f) Use $t = (\bar{Y}_{1..} - \bar{Y}_{2..}) \Big/ \sqrt{\frac{2}{9} E(MS(B(A)))} = \underbrace{(1.864 - 6.091)}_{-4.227} \Big/ \sqrt{\frac{2}{9}(.366)} = 14.822$

follows a t with 4 df and is significant for even very small α. Thus conclude that $\alpha_1 \neq \alpha_2$.

g) The improvements for the two drugs (averaged over batches and individuals) is not the same for both drugs.

(Note: if one had the opportunity to look at the ANOVA table, one would not conclude that there is variability from batch to batch, leading to a possibly stronger conclusion)

improvement

Solution (continued)

Results for: improvement-MTW

ANOVA: improvement versus drug, batch

Factor	Type	Levels	Values
drug	fixed	2	1 2
batch(drug)	random	3	1 2 3

Analysis of Variance for improvem

Source	DF	SS	MS	F	P
drug	1	80.400	80.400	219.44	0.000
batch(drug)	4	1.466	0.366	1.38	0.298
Error	12	3.181	0.265		
Total	17	85.046			

Source		Variance component	Error term	Expected Mean Square for Each Term (using unrestricted model)
1	drug		2	(3) + 3(2) + Q[1]
2	batch(drug)	0.03377	3	(3) + 3(2)
3	Error	0.26509		(3)

Results for: improvement.MTW

Data Display

obs.	drug	batch	improvement	
1	1	1	1.257	
2	1	1	1.415	
3	1	1	2.172	
4	1	2	2.743	
5	1	2	2.250	*
6	1	2	2.179	
7	1	3	1.000	
8	1	3	1.657	
9	1	3	2.107	
10	2	1	6.007	
11	2	1	6.457	
12	2	1	5.329	
13	2	2	5.936	
14	2	2	6.493	
15	2	2	5.693	
16	2	3	6.857	
17	2	3	5.550	
18	2	3	6.500	