MASSACHUSETTS INSTITUTE OF TECHNOLOGY
Department of Civil and Environmental Engineering

1.017/1.010 Computing and Data Analysis for Environmental Applications / Uncertainty in Engineering

Problem Set 9: Two-way ANOVA and Regression (Solutions provided at end of each problem)
Due: Tuesday, Dec. 2, 2003

Please turn in a hard copy of your MATLAB program as well as all printed outputs (tables, plots, etc.) required to solve the problem.

Problem 1

Continue your investigation of Massachusetts Water Resources Authority (MWRA) Boston Harbor data by carrying out a two factor ANOVA which considers both rainfall and date. For Factor A (rainfall) your treatments should be the same three rainfall intensity categories used in Problem Set 8. For Factor B (date) use two treatments: 1) dates through 1991 and 2) dates after 1991. The dates are included in the file lowercharles.txt that you used in Problem Set 8 .

You should use the internal MATLAB function `ANOVA2` to carry out your two-factor analysis. This function can handle treatments that have different numbers of replicates. However, if you prefer to adopt the equal replicate assumption, you can force each treatment to have the same number of replicates (as assumed in the lecture notes) by randomly selecting the same number of measurements from the coliform population available for each treatment combination.

As before, convert the coliform count *C* to the transformed count $C_T = ln\ (C+1)$ and check for normality. Present the two factor ANOVA table, report the *p* value, and discuss the significance of your results.

Problem 1 Solution

```
% Problem Set 9 -- Problem 1
clear all
close all
% edited lowercharles3 in excel: deleted unnecessary columns
% replaced blank rain with zeros
load lowercharles3.txt;
date=lowercharles3(:,1);
coli=log(lowercharles3(:,2)+1);
rain=lowercharles3(:,3);
```

```
p=1;
q=1;
r=1;
s=1;
u=1;
v=1;
for i=1:length(coli)
    if rain(i)==0&date(i)<33790
        T1(p)=coli(i);
        p=p+1;
    elseif rain(i)==0&date(i)>=33790
        T4(s)=coli(i);
        s=s+1;
    elseif (rain(i)>0&rain(i)<=0.25)&date(i)<33790
        T2(q)=coli(i);
        q=q+1;
    elseif (rain(i)>0&rain(i)<=0.25)&date(i)>=33790
        T5(u)=coli(i);
        u=u+1;
    elseif rain(i)>.25&date(i)<33790
        T3(r)=coli(i);
        r=r+1;
    elseif rain(i)>.25&date(i)>=33790
        T6(v)=coli(i);
        v=v+1;
    end
end
number=min([(length(T1)), (length(T2)),
(length(T3)),(length(T4)), (length(T5)), (length(T6))]);
index1=randperm(length(T1));
ctreat1=T1(index1(1:number));
index2=randperm(length(T2));
ctreat2=T2(index2(1:number));
index3=randperm(length(T3));
ctreat3=T3(index3(1:number));
index4=randperm(length(T4));
ctreat4=T4(index4(1:number));
index5=randperm(length(T5));
ctreat5=T5(index5(1:number));
index6=randperm(length(T6));
ctreat6=T6(index6(1:number));
figure
normplot([ctreat1-mean(ctreat1),ctreat2-mean(ctreat2),ctreat3-
mean(ctreat3),ctreat4-mean(ctreat4), ctreat5-mean(ctreat5),
ctreat6-mean(ctreat6)])
title('Normality check for transformed coliform data')
matrix=[ctreat1', ctreat2', ctreat3', ctreat4', ctreat5',
ctreat6'];
```

```
anova2(matrix,58)
```

Problem 2

Consider the following set of data relating evaporation rate L (mm/hr) to soil temperature T (in degree C.):

Temperature (T_i): [10 12 15 17 22 24]

Evaporation rate (R_i): [5.0001 5.0464 10.7850 2.8780 12.2845 14.6869]

Carry out a regression analysis for this problem using the following model:

$$\boldsymbol{L} = a_1 T + \boldsymbol{e} \qquad [1]$$

$$E[\boldsymbol{L}] = a_1 T \qquad [2]$$

This model implies that the measurement $\boldsymbol{L}_i$ at temperature T_i is a random variable composed of a term proportional to T_i and a zero mean random residual $\boldsymbol{e}_i$:

$$\boldsymbol{L}_i = a_1 T_i + \boldsymbol{e}_i \quad ; \quad i = 1,\ldots,6 \qquad [3]$$

Carry out the regression analysis by performing the following steps:

1. Write a symbolic expression for *SSE*, the sum-of-squared prediction errors, as a function of $L_{i,}$ and T_i (don't substitute numerical values in for L_i and T_i at this point).

2. Find the value of a_1 that minimizes the *SSE* by setting the derivative d*SSE*/da_1 equal to zero. Solve the resulting minimization equation for a_1. This minimizing value of a_1 is the least-squares estimate $\hat{\boldsymbol{a}}_1$. Your symbolic expression for $\hat{\boldsymbol{a}}_1$ should be a linear function of the individual measurements $\boldsymbol{L}_1 \ldots \boldsymbol{L}_6$.

3. Substitute the numerical values for the T_i and $\boldsymbol{L}_i$ ($i = 1,\ldots 6$) in your symbolic expression for $\hat{\boldsymbol{a}}_1$. Then substitute $\hat{a}_1$ in the prediction equation $\hat{L} = \hat{a}_1 T$ to get an expression that predicts L for any T.

4. Use MATLAB to plot your prediction equation over an appropriate range of T values. Also, use `hold on` to plot the specified values of T_i and measured values of L_i on the same axes. Comment on the quality of your "fit".

5. Compute the mean and variance of $\hat{\boldsymbol{a}}_1$ from your estimation equation. To do this you need to substitute the right-hand side of Eq [3] above for each of the measurements $\boldsymbol{L}_i$, $i = 1\ldots 6$ in the $\hat{\boldsymbol{a}}_1$ equation. Then invoke the definitions of the mean and variance and

take advantage of the properties of linear functions of independent random variables (in this case the independent random variables are the $\boldsymbol{e}_i$ ’s).

6. Assume the residual errors are zero mean and normally distributed. Compute a two-sided 95% confidence interval for a_1 using the appropriate small sample (t) statistic.

7. Finally, check all of your results by analyzing the same data with the internal MATLAB function `regress`. Please carefully review the MATLAB help documentation for this function to make sure that the confidence interval computed by MATLAB and the confidence interval you derive yourself are consistent.

Problem 2 Solution

```
% Problem Set 9 -- Problem 2
clear all
close all
T=[10 12 15 17 22 24];                              %[deg C]
L=[5.0001 5.0464 10.7850 2.8780 12.2845 14.6869];   %[mm/hr]
N=6
Y=L';
H=[T'];
ahat=(H'*H)\(H'*Y)
figure
plot(T,L,'*')
hold
x=5:.01:25;
plot(x,ahat*x)
[B,BINT,R,RINT,STATS] =regress(Y,H,.05)
vare=1/(6-1)*sum((L-ahat*T).^2)
% confidence interval limits
upperlim=ahat-tinv(.025,N-1)*sqrt(vare*inv(H'*H))
lowerlim=ahat-tinv(.975,N-1)*sqrt(vare*inv(H'*H))
```

Problem 3

In this problem you will perform a regression analysis of temporal trends in global population data. The data you need are located at the UN Population Division web site (Panel 2: Detailed Data):

> http://esa.un.org/unpp/index.asp?panel=2
>
> Download the following annual data sets:
>
> > Population by sex (annual), medium variant
> > Use the column labeled (“both sexes combined”)

For the following 2 regions:

More developed regions
Less developed regions

For all years 1950-2000.

You can examine the data using the “display” option on the web site but should download the data as a .CSV file. If you have EXCEL you can view the downloaded data in a spreadsheet and then extract particular sections to put into MATLAB.

Carry out a regression analysis for each of the 2 data sets using a MATLAB code that constructs the various arrays and plots required to analyze the data and display the results. You can use the MATLAB function `regress` in your code (this function will compute the regression coefficients for you). Note that the matrix *X* mentioned in the `regress` documentation is the same as the matrix *H* discussed in the notes.

Your regression analysis should fit a quadratic function to the 1950-2000 data and, for each of the 2 data sets, plot the regression function and the data over the period 1950-2030 (the portion from 2001-2030 will include only the regression function since it predicts beyond the data period). To improve numerical accuracy define your independent variable to be x = *year*-1949, so x varies from 1 to 81.

Plot prediction confidence intervals over the entire 1950-2030 period to get a feeling for how well your function might predict population. Use the prediction confidence interval expression given at the end of the Class 23 Lecture Notes. Note that you will need to compute the sample standard deviation s_e of the measurement residuals (these residuals are returned by `regress` in the array `r`). You will also have to compute the matrix $(H'H)^{-1}$, which is just `inv(H*H')` in MATLAB. In order to construct your confidence interval curves you should evaluate the x value in the array $h(x)$ for each year from 1950 to 2030 [i.e. $h(1) = a_1+a_2(1)+a_3 (1)^2$, $h(2) = a_1+a_2(2)+a_3 (2)^2$, etc.].

Comment on the implications of your predictions. Also, comment on the limitations of the prediction confidence intervals produced by your regression analysis.

Problem 3 Solution

```
% Problem Set 9 -- Problem 3
clear all
close all
load UNPop.txt
year=UNPop(:,1)-1949;
morepop=UNPop(:,2);
lesspop=UNPop(:,3);
H=[ones(length(year),1),year,year.*year];
Ymore=morepop;
```

```
Yless=lesspop;
ahatmore=(H'*H)\(H'*Ymore)
ahatless=(H'*H)\(H'*Yless)
[Bm,BINTm,Rm,RINTm,STATSm] =regress(Ymore,H,.05)
[Bl,BINTl,Rl,RINTl,STATSl] =regress(Yless,H,.05)
x=1:1:81;
sem=std(Rm)
sel=std(Rl)
mat=inv(H'*H);
smallh=[ones(length(x),1),x',x'.^2];
yhatmore=ahatmore(1)+ahatmore(2)*x+ahatmore(3)*x.^2;
yhatless=ahatless(1)+ahatless(2)*x+ahatless(3)*x.^2;
for j=1:length(x)
     ymoreupper(j)=yhatmore(j)+ …
     1.96*(smallh(j,:)*sem^2*mat*smallh(j,:)')^(1/2);
     ymorelower(j)=yhatmore(j)- …
     1.96*(smallh(j,:)*sem^2*mat*smallh(j,:)')^(1/2);
     ylessupper(j)=yhatless(j)+ …
     1.96*(smallh(j,:)*sel^2*mat*smallh(j,:)')^(1/2);
     ylesslower(j)=yhatless(j)- …
     1.96*(smallh(j,:)*sel^2*mat*smallh(j,:)')^(1/2);
end
% for more developed regions: calculate upper and
% lower curves:
figure (1)
plot(year+1949,morepop,'*')
hold
plot(x+1949,yhatmore)
title('Population in More Developed Countries vs Year')
xlabel('Year')
ylabel('Population')
plot(x+1949,ymoreupper,'--')
plot(x+1949,ymorelower,'--')
% for less developed regions: calculate upper and
% lower curves:
figure (2)
plot(year+1949,lesspop,'*')
hold
plot(x+1949,yhatless)
title('Population in Less Developed Countries vs Year')
xlabel('Year')
ylabel('Population')
plot(x+1949,ylessupper,'--')
plot(x+1949,ylesslower,'--')
```