THE DESIGN AND MIXED-MODEL ANALYSIS OF EXPERIMENTS

PRACTICAL APPENDIX A

GENSTAT 5 USAGE

The answers to questions 1–4 were given in the lecture notes or provided with the question.

5. Generating a Voter sample

Genstat's random number generator will be used to investigate the sampling distribution of the F statistic in the case of the Voter turnout data from example II.2. In this example, the regression of voter turnout on the % advertising expenditure was examined. Suppose that we assume that the population model underlying this data is actually

$$E[Y] = 45 \text{ and } \sigma^2 = 10$$

with Y normally distributed. This model says that the turnout is constant and displays variation such that the standard deviation in the turnouts is $\sqrt{10} = 3.1623$. Our model assumes that in the population the slope is zero.

The following commands will generate a sample of 20 observations from such a population and then analyse data, assuming the sample has the same values of the explanatory variable as that supplied in the lecture. First we need the expenditure values that have been observed. The data has been saved in the Excel spreadsheet file Voter.xls in the directory G:\Disciplina\Genstat. Use *File > Open* with the *All Files* option selected in the *Files of Type* list box. Secondly the commands have been save in Votergen1.gen. Open this file and run the commands it contains. Give it a six-digit number for the seed when it asks for one.

```
"load voter.xls from g:\disciplina\genstat so expend available"
  question {PREAMBLE=!T('What 6-digit number do you wish to use to \
seed the random number generator?'); \
    MODE=v; TYPE=scalar; PRESENT=yes; UPPER=999999; RESPONSE=_seed]
  print _seed
"generate turnout according to the model"
  calc turnout=45+ned(urand(_seed; 20))*sqrt(10)
  print turnout,expend
  describe [selection=nobs,mean,sd,var,min,max] turnout
"analyze generated data"
  model turnout
  terms expend
  fit [fprob=y] expend
  rkeep deviance=res_ss; df=res_df
  calc f=(variance(turnout)*19-res_ss)/(res_ss/res_df)
```

```
  &    p=1-fratio(f; 1; 18)
"check calculations"
  print res_ss,res_df,f,p
```

The output produced by these commands is as follows:

```
Genstat 5  Release 4.1  (PC/Windows NT)              19 March 2000 14:54:42
Copyright 1998, Lawes Agricultural Trust (Rothamsted Experimental Station)

                ________________________________________________

                Genstat 5 Fourth Edition - (for Windows)
                Genstat 5 Procedure Library Release PL11
                ________________________________________________

   3  "Data taken from File: D:/ANALYSES/LM/VOTER.XLS"
   4  DELETE [redefine=yes] turnout,expend
   5  VARIATE [nvalues=20] turnout
   6  READ turnout

    Identifier   Minimum      Mean   Maximum    Values   Missing
       turnout     24.60     45.71     64.80        20         0

   9  VARIATE [nvalues=20] expend
  10  READ expend

    Identifier   Minimum      Mean   Maximum    Values   Missing
        expend     23.00     36.99     50.10        20         0

  13
  14  "load voter.xls from g:\disciplina\genstat so expend available"
  15    question {PREAMBLE=!T('What 6-digit number do you wish to use to \
 -16  seed the random number generator?'); \
  17      MODE=v; TYPE=scalar; PRESENT=yes; UPPER=999999; RESPONSE=_seed]
  18    print _seed


      123456

  19  "generate turnout according to the model"
  20    calc turnout=45+ned(urand(_seed; 20))*sqrt(10)
  21    print turnout,expend

     turnout      expend
       44.39       28.50
       45.19       48.30
       46.30       40.20
       50.81       34.80
       46.07       50.10
       45.44       44.00
       42.24       27.20
       45.72       37.80
       51.21       27.20
       47.11       46.10
       43.70       31.30
       44.12       50.10
       40.44       31.30
       46.93       24.80
       44.91       42.20
       43.56       23.00
       49.73       30.10
       41.75       36.50
       47.35       40.20
       46.23       46.10
```

```
  22    describe [selection=nobs,mean,sd,var,min,max] turnout

 Summary statistics for turnout

      Number of observations = 20
                        Mean = 45.660
                     Minimum = 40.440
                     Maximum = 51.213
                    Variance = 7.831
          Standard deviation = 2.798
  23  "analyze generated data"
  24    model turnout
  25    terms expend
  26    fit [fprob=y] expend


26.............................................................


***** Regression Analysis *****

 Response variate: turnout
     Fitted terms: Constant, expend


*** Summary of analysis ***

              d.f.         s.s.         m.s.      v.r.  F pr.
Regression       1          0.0        0.039      0.00  0.946
Residual        18        148.7        8.264
Total           19        148.8        7.831

Residual variance exceeds variance of response variate
Standard error of observations is estimated to be 2.87
* MESSAGE: The following units have large standardized residuals:
         Unit     Response    Residual
            9        51.21        2.07


*** Estimates of parameters ***

                  estimate         s.e.     t(18)
Constant             45.47         2.86     15.92
expend              0.0051       0.0752      0.07

  27    rkeep deviance=res_ss; df=res_df
  28    calc f=(variance(turnout)*19-res_ss)/(res_ss/res_df)
  29    &    p=1-fratio(f; 1; 18)
  30  "check calculations"
  31    print res_ss,res_df,f,p

      res_ss      res_df           f           p
       148.7       18.00    0.004663      0.9463
```

The p value contain in p is the same as in the column F pr. in the ANOVA table.

If you were to take a large number of samples of this kind and compute the F statistic and p value for each one, what distribution would expect to obtain for the F statistic? What proportion of F statistics would you expect to have a p value of 0.05 or less?

One would expect an F distribution with one and 18 degrees of freedom.

I would expect 0.05 or 5% of F statistics to have a p value of 0.05 or less since the p value is the probability of an F statistic equal to or larger than the one for

which the p value is calculated. By definition, only 5% can have a p value of 0.05 or smaller.

6. Generating 250 Voter samples

Use Genstat to generate 250 samples of the kind generated in exercise 5. A program for doing this is listed below and is available in the file votegen1.gen in the directory g:\discplina\genstat.

Open the file in Genstat and run it. Give it a six-digit number for the seed when it asks for one. Does the distribution look right? What about the number of F statistics with a p below 0.05? What do you conclude about the value of 182.26 for F for this data set which was given in the lecture?

```
"load voter.xls from g:\disciplina\genstat so expend available"
  question {PREAMBLE=!T('What 6-digit number do you wish to use to \
seed the random number generator?'); \
    MODE=v; TYPE=scalar; PRESENT=yes; UPPER=999999; RESPONSE=_seed]
  print _seed
  calc turnout=urand(_seed; 20)
"generate turnout according to the model for 250 samples"
  vari [250] f
  for i=1...250
    calc turnout=45+ned(urand(0; 20))*sqrt(10)
    model turnout
    terms expend
    fit [print=*] expend
    rkeep deviance=res_ss; df=res_df
    calc f$[i]=(variance(turnout)*19-res_ss)/(res_ss/res_df)
  endfor
"examine computed Fs"
  describe [selection=nobs,min,q1,q3,max] f
  histogram f
  calc p=1-fratio(f; 1; 18)
  & signif=sum(p <= 0.05)
  print signif
```

The output produced by these commands is as follows:

```
Genstat 5  Release 4.1  (PC/Windows NT)              19 March 2000 15:04:40
Copyright 1998, Lawes Agricultural Trust (Rothamsted Experimental Station)

                 ________________________________________________

                 Genstat 5 Fourth Edition - (for Windows)
                 Genstat 5 Procedure Library Release PL11
                 ________________________________________________

   3  "Data taken from File: D:/ANALYSES/LM/VOTER.XLS"
   4  DELETE [redefine=yes] turnout,expend
   5  VARIATE [nvalues=20] turnout
   6  READ turnout

    Identifier   Minimum      Mean   Maximum    Values   Missing
       turnout     24.60     45.71     64.80        20         0

   9  VARIATE [nvalues=20] expend
  10  READ expend

    Identifier   Minimum      Mean   Maximum    Values   Missing
        expend     23.00     36.99     50.10        20         0

  13
```

```
  14  "load voter.xls from g:\disciplina\genstat so expend available"
  15    question {PREAMBLE=!T('What 6-digit number do you wish to use to \
 -16  seed the random number generator?'); \
  17      MODE=v; TYPE=scalar; PRESENT=yes; UPPER=999999; RESPONSE=_seed]
  18    print _seed


      123456

  19    calc turnout=urand(_seed; 20)
  20  "generate turnout according to the model for 250 samples"
  21    vari [250] f
  22    for i=1...250
  23      calc turnout=45+ned(urand(0; 20))*sqrt(10)
  24      model turnout
  25      terms expend
  26      fit [print=*] expend
  27      rkeep deviance=res_ss; df=res_df
  28      calc f$[i]=(variance(turnout)*19-res_ss)/(res_ss/res_df)
  29    endfor
  30  "examine computed Fs"
  31    describe [selection=nobs,min,q1,q3,max] f

 Summary statistics for f

      Number of observations = 250
                     Minimum = 0.000
                     Maximum = 13.963
             Lower quartile = 0.067
             Upper quartile = 1.331
  32    histogram f

Histogram of f


              -  1.5  194 **************************************************
          1.5 -  3.0   33 ********
          3.0 -  4.5    9 **
          4.5 -  6.0    7 **
          6.0 -  7.5    2 *
          7.5 -  9.0    2 *
          9.0 - 10.5    2 *
         10.5 - 12.0    0
         12.0 - 13.5    0
         13.5 -         1

 Scale:  1 asterisk represents 4 units.

  33    calc p=1-fratio(f; 1; 18)
  34    & signif=sum(p <= 0.05)
  35    print signif

      signif
       14.00
```

The distribution does look like an F distribution with one and 18 degrees of freedom. For these numbers of degrees of freedom, the distribution does not form a peak but asymptotes with the vertical axis.

There were 14 out of 250 F statistics with a p value less than 0.05. This equals 5.6% which is quite close to the 5% expected.

The value of 182.26 given in the lecture notes is clearly a very unusual value to get when the slope is zero.