Homework 2

IGP 304: Statistics for Biomedical Research (Spring 2007)
due on February 12, 2007

The following questions relate to the *10.7.ERpolymorphism.dta* and *2.12.Poison.dta* data sets, which can be found in "Statistical Modeling for Biomedical Researchers", 2nd Ed., *in press*, by William Dupont. The data set can be downloaded from the web

http://biostat.mc.vanderbilt.edu/dupontwd/wddtext/ or the homework section of the course web.

- Stata commends useful for this homework can be found in Stata notes for classes on the course web.

1. (25 points total) Use *10.7.ERpolymorphism.dta* which was obtained by Parl et al. (1989) who studied the relationship between age at breast cancer diagnosis and estrogen receptor (ER) genotype. Answer to the following questions.

   (a) (3 points) Summarize age by genotype using:

   ```
   summarize age if genotype==1
   summarize age if genotype==2
   summarize age if genotype==3
   ```

   Assuming normal distribution, calculate 95% confidence interval for the mean of age by genotype using this information (**Show your work.**).

   (b) (2 points) Draw box plots for age by genotype on a single graph. Draw dot plots for age by genotype on a single graph.

   (c) Do a one-way analysis of variance to test the null hypothesis that the age at diagnosis does not vary with genotype.

   - (3 points) What are the sum of squares (SS), mean squares (MS) and degree of freedom (df) for between-groups and total? If you know the SS, MS and df for between-groups and total, can you calculate the SS, MS and df for within-groups?
   - (3 points) (i) Show how F-value can be calculated from the SS; (ii) report *P*-value; and (iii) what would you conclude from the result based on the hypothesis test?
   - (2 point) What assumptions did you make to do a one-way analysis of variance and do you think the assumptions are appropriate?

   (d) Repeat analysis using linear regression.

   - (5 points) To use linear regression, you need to generate two dummy variables. Show the output. Compare the upper panel of the output with the ANOVA table from (c). What do you observe?

- (2 points) What is the pooled standard deviation?
- (5 points) Interpret the estimates of parameters.

2. (15 points total) Use *2.12.Poison.dta*. This data set was obtained by Brent et al. (1999) who measured baseline plasma glycolate and arterial pH on 18 patients admitted for ethylene glycol poisoning.

   (a) (3 points) Regress plasma glycolate against arterial pH. Draw a scatter plot of plasma glycolate against arterial pH together with the estimated linear regression line. What is the best fitted line?

   (b) (3 points) What is the slope estimate and the 95% confidence intervals for the slope estimate? Interpret the slope estimate.

   (c) (2 points) State your conclusion based on the hypothesis testing and the confidence intervals.

   (d) (3 points) What is the estimated average plasma glycolate if a patient has arterial pH of 7.1?

   (e) (2 points) What proportion of the total variation in plasma glycolate levels is explained by this regression?

   (f) (2 points) Draw a scatter plot of plasma glycolate against arterial pH together with the estimated linear regression line and the 95% confidence intervals band.

3. (5 points) Below find a partial output from a regression analysis. Based on the information, fill in the following ANOVA table.

```
. regress y x
      Source |       SS       df       MS              Number of obs =     636
-------------+------------------------------           F(  1,   634) =
       Model |                 1                       Prob > F      =  0.0000
    Residual |                                         R-squared     =  0.3863
-------------+------------------------------           Adj R-squared =  0.3854
       Total |   58.858423                             Root MSE      =  .23868

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .0299317   .0014982    19.98   0.000     .0269897    .0328737
       _cons |  -2.118609   .1861004   -11.38   0.000    -2.484057   -1.753161
------------------------------------------------------------------------------
```

Source of Variability	Sum of Squares	df	Mean Square	F
Regression				
Error (residual)				
Total				

4. (5 points total) **Show your work.** Below find a table which summarizes means and standard deviations of BMI for males and females. Assuming a normal distribution, calculate separately for males and females the probability of a BMI < 25, between $25 - 29.9$, and 30 or more kg/m^2. The probabilities can be obtained from the table for the standard normal distribution(Table A1 in [ES]) or by using the following `Stata` commend. For example, you can use the following commend to calculate 1-P(Z < 2): `display 1-normprob(2)`

```
Men:
    Variable |       Obs        Mean       Std. Dev.
-------------+----------------------------------
         BMI |       100      25.24071    5.204428

Women:
    Variable |       Obs         Mean       Std. Dev.
-------------+----------------------------------
         BMI |       218      24.79305    6.086048
```

5. (10 points total) **Show your work next to your response.** Below find a table from a hypothetical study. This study compares treatment and control groups in a randomized trial. Several measurements were made before ("pretest") and after ("posttest") the treatment. The table shows a part of this study.

		Mean (SD)			
Variable	n	Pretest	Posttest	Improvement	*P*-value
Measurement 1					
Treatment	22	5.0 (2.8)	2.9 (2.2)	2.1 (3.1)	0.02
Control	20	5.2 (2.1)	4.3 (2.2)	0.9 (2.8)	0.16
Measurement 2					
Treatment	33	161.6 (70.4)	187.4 (68.8)	25.8 (41.4)	0.009
Control	29	183.9 (69.5)	190.5 (68.2)	6.6 (41.1)	0.37

(a) (4 points) The average change in Measurement 2 for the treatment and control groups are 25.8 (41.4) and 6.6 (41.1), respectively. A valid 95% confidence interval for the change in the outcome score for the treatment group is:

( ) *a.* no information provided
( ) *b.* -15.6 to 67.2
( ) *c.* 20.0 to 31.5
( ) *d.* 14.1 to 37.5
( ) *e.* 11.1 to 40.5

(b) Use the Stata log to answer to the following two questions.

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
         |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       x |      33        25.8    7.206815        41.4     11.1202     40.4798
       y |      29         6.6    7.632078        41.1   -9.033604     22.2336
---------+--------------------------------------------------------------------
combined |      62    16.81935    5.339721    42.04501    6.141922    27.49679
---------+--------------------------------------------------------------------
    diff |                19.2      10.502               -1.807125    40.20712
------------------------------------------------------------------------------
    diff = mean(x) - mean(y)
Ho: diff = 0                                     degrees of freedom =       60

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9638         Pr(|T| > |t|) = 0.0725          Pr(T > t) = 0.0362
```

i. (3 points) Suppose the investigator is interested in a one-sided test that the average change is the same (or worse) in the two groups versus better in the treatment group. The corresponding *P*-value is:

( ) *a.* 0.0725
( ) *b.* 0.0021
( ) *c.* 0.4810
( ) *d.* 0.0362
( ) *e.* 0.3945

What would the investigator conclude?

ii. (3 points) The *t* statistic for testing a two-sided test of the null hypothesis that the treatment and control populations have the same mean change would be:

( ) *a.* 19.2
( ) *b.* 1.83
( ) *c.* 10.4
( ) *d.* 3.58
( ) *e.* no information available

What would the investigator conclude?