Statistics 515 - Fall 2003 - Practice Exam 3 *(based on past exams)*

Part I: Answer three of the following four questions. If you complete more than three, I will grade only the first three. Five points each.

1) Define what is meant by the *p-value* (or the *observed significance level*) of a test. __________

____

____.

2) (Circle the correct answers) When conducting a two-sample t-test that two population means are equal we use **s_1^2 and s_2^2 / s_p^2** and **n_1+n_2-2 degrees of freedom / Satterthwaite's formula** if the variances are equal.

3) (Circle the correct answers) A student achieved a score of 780 out of 800 on an aptitude test. If you knew nothing else about the student, regression to the mean would imply that they would score **lower / about the same / higher** if they retook the test. A student scoring a 500 out of 800 (near the average) would score **lower / about the same / higher** if they retook it.

4) Say we reject the null hypothesis β_1=0 in a regression problem. Briefly explain why can't we automatically assume that changing *x* causes *y* to change?

Part II: Answer every part of the next three problems. Read each problem carefully, and show your work for full credit. Twenty points each.

1) A candidate for political office wants to determine if there is a difference in his popularity between men and women. To test this he collects a sample of 250 men and 250 women and records how many of them plan on voting for him in the upcoming election.

a) State the appropriate null and alternate hypothesis for determining whether the candidate differs in popularity between men and women. Be sure to identify what the using mean in terms of the problem (e.g. if you use $\mu, p, \sigma^2, s^2, \bar{x}, \hat{p}$ say what parameter(s) you are the symbol stands for.)

b) Of those sampled, 105 of the men and 128 of the women plan on voting for the candidate. Report the p-value for the test of hypothesis in A.

c) Besides the sample being randomly chosen, what other assumption(s) are required to trust the test in part B? If possible, check that the assumption(s) hold.

2) The following partial ANOVA table is for comparing how many hours it takes for different headache remedies to provide relief.   To test this a group of patients was randomly divided into separate groups, with each group taking a different remedy.

Source	SS	DF	MS	F	p-value
Treatments	3.3054	__	______	______	0.000
Error	1.9553	30	______		
Total	______	32			

a) Complete the above ANOVA table by filling in the missing values.

b) How many different headache remedies were compared in this experiment? ___________
How many total patients were used in this experiment? ____________

c) What null and alternate hypothesis are being tested by the p-value in this ANOVA table? (Be sure to identify any parameters that you use?)

d) Do the headache remedies provide the same average relief, or is there a difference? (How did you decide?)

3) The attached data set concerns the eruptions of the Old Faithful Geyser from August 1 to August 8, 1978. `date` is the day of the eruption, `interval` is the length of time until the next eruption (in minutes), and `duration` is the length of the last eruption (in minutes).   The goal is to predict the time until the next eruption (the interval) from the length of the last eruption (the duration).

a) It is possible to check three of the four regression assumptions by using the graphs that are produced by SAS. Say which three assumptions those are and why they seem to be met in this case.

b) Assuming the assumptions of the regression model are met, what is the p-value for testing the hypothesis that $\beta_1 = 0$?   Do we accept or reject this null hypothesis at $\alpha = 0.01$? Does the duration of the previous eruption help predict the time until the next eruption?

c) If old faithful just erupted for 3 minutes, how long do you predict it will be until the next eruption?

d) What is the estimate of the standard deviation of the errors (σ) for this regression?

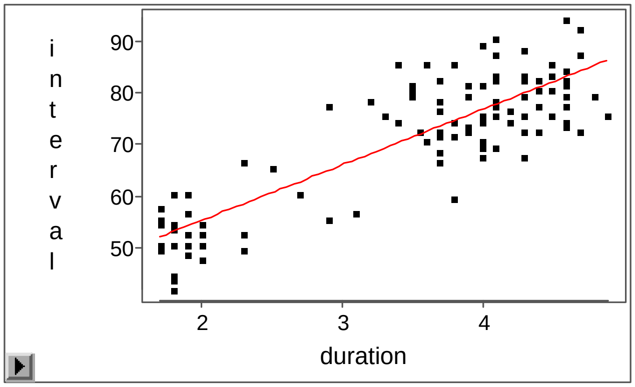
e) What percent of the variation or error in predicting the time until the next eruption is explained by the duration of the previous eruption?

```
DATA oldfaith;
INPUT date $ interval duration;
CARDS;
1       78      4.4
1       74      3.9
1       68      4.0
1       76      4.0
1       80      3.5
1       84      4.1
1       50      2.3
1       93      4.7
1       55      1.7
1       76      4.9
1       58      1.7
1       74      4.6
1       75      3.4
2       80      4.3
2       56      1.7
2       80      3.9
2       69      3.7
2       57      3.1
2       90      4.0
2       42      1.8
2       91      4.1
2       51      1.8
2       79      3.2
2       53      1.9
2       82      4.6
2       51      2.0
3       76      4.5
3       82      3.9
3       84      4.3
3       53      2.3
3       86      3.8
3       51      1.9
3       85      4.6
3       45      1.8
```

```
3       88      4.7
3       51      1.8
3       80      4.6
3       49      1.9
3       82      3.5
4       75      4.0
4       73      3.7
4       67      3.7
4       68      4.3
4       86      3.6
4       72      3.8
4       75      3.8
4       75      3.8
4       66      2.5
4       84      4.5
4       70      4.1
4       79      3.7
4       60      3.8
4       86      3.4
5       71      4.0
5       67      2.3
5       81      4.4
5       76      4.1
5       83      4.3
5       76      3.3
5       55      2.0
5       73      4.3
5       56      2.9
5       83      4.6
5       57      1.9
5       71      3.6
5       72      3.7
5       77      3.7
6       55      1.8
6       75      4.6
6       73      3.56
6      70      4.0
```

```
6       83      3.7
6       50      1.7
6       95      4.6
6       51      1.7
6       82      4.0
6       54      1.8
6       83      4.4
6       51      1.9
6       80      4.6
6       78      2.9
7       81      3.5
7       53      2.0
7       89      4.3
7       44      1.8
7       78      4.1
7       61      1.8
7       73      4.7
7       75      4.2
7       73      3.9
7       76      4.3
7       55      1.8
7       86      4.5
7       48      2.0
8       77      4.2
8       73      4.4
8       70      4.1
8       88      4.1
8       75      4.0
8       83      4.1
8       61      2.7
8       78      4.6
8       61      1.9
8       81      4.5
8       51      2.0
8       80      4.8
8       79      4.1
;
```

```
PROC INSIGHT;
OPEN oldfaith;
FIT interval=duration;
RUN;
```



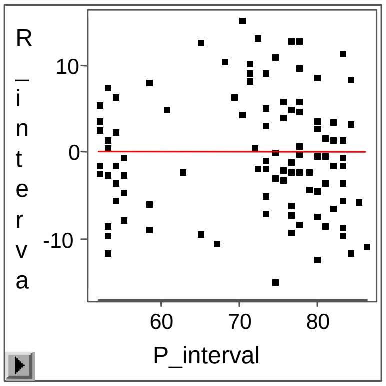
```
                         Model Equation
interval         =      33.8212      +      10.7412      duration
```

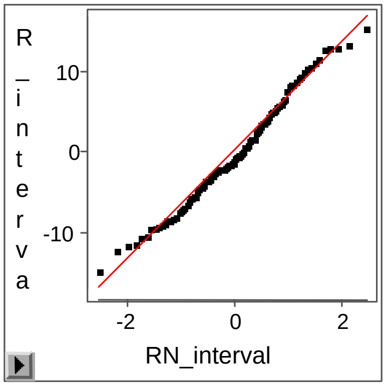
```
                          Summary of Fit
Mean of Response          71.0000        R-Square        0.7370
Root MSE                   6.6815        Adj R-Sq        0.7345
```

```
                                   Analysis of Variance
 Source      DF      Sum of Squares      Mean Square      F Stat       Pr > F
Model         1          13134.6089       13134.6089      294.22       <.0001
Error       105           4687.3911          44.6418
C Total     106          17822.0000
```

```
                                               Parameter Estimates
 Variable        DF       Estimate       Std Error         t Stat         Pr >|t|        Tolerance        Var Inflation
Intercept         1       33.8212        2.2617            14.95          <.0001             .                       0
duration          1       10.7412        0.6262            17.15          <.0001          1.0000                  1.0000
```

```
                     95% C.I. for Parameters
 Variable          Estimate            Lower              Upper
Intercept          33.8212            29.3367            38.3057
duration           10.7412             9.4996            11.9829
```


R_interva
P_interval


R_interva
RN_interval