Spring 2000 STAT 5303 Sample Midterm

Instructions:

- Write your name on your exam book. Write your name on your formula sheet. You will need to turn both in. Please keep your copy of this exam because it will be beneficial to have it on hand when you look over your graded work.
- Work independently. If you would like me to rephrase any questions, just ask.
- Answer all any 4 of the 5 questions.
- Each question is worth the same number of points. Sub-parts within each problem will each have about the same number of points.

1. The ability of ecologists to identify regions of greatest species richness could have an impact on the preservation of genetic diversity, a major objective of the World Conservation Strategy. The article "Prediction of Rarities from Habitat Variables: Coastal Plain Plants on Nova Scotian Lakeshores" (*Ecology* (1992): 1852-1859) used a sample of 37 lakes to obtain the estimated regression equation $\hat{y} = 3.89 + .033x_1 + .024x_2 + .023x_3 - .0080x_4 - .13x_g - .72x_g$ where y=species richness, x_1=watershed area, x_2=shore width, x_3=poor drainage(%), x_4=water color (total color units), x_5 = sand(%) and x_6=alkalinity. The value for R^2 was reported as 0.83.
   a. Use a test with significance level 0.01 to decide whether at least one of the betas is nonzero. [Note: an alternative form of the F statistic equation is $F = \dfrac{R^2 / k}{(1-R^2)/(n-(k+1))}$, where k=number of explanatory variables and n=number of observations.]
   b. In what situation(s) would one want to use Adjusted R^2 in place of R^2.
   c. Predict the species richness for a region with watershed area of 1800, shore width 565, 15% poor drainage, water color 4, 34% sand and alkalinity level 8.
   d. Explain why two predictions, both with values for each explanatory variable value well within the "typical" range could have vastly differing prediction intervals, one **much** wider than the other. [Note: this won't necessarily happen with the dataset used in this problem.]

2. Suppose we have a categorical predictor along with the other continuous explanatory variables that we believe to be possibly related to the response variable, y.
   a. Explain how we could include this categorical predictor in our model.
   b. Explain how we could test the hypothesis that this categorical predictor had no predictive value (versus the alternative that it does)?

3. An experiment was performed to investigate the relationship between x1, x2 and y. The researchers, however, did not know the exact relationship between inputs (x1 and x2) and response (y). Thus, they fit a "full second order model" which includes quadratic terms for x1 and x2 and the interaction between x1 and x2. Such a model would allow any quadratic-like curvature between inputs and response to be estimated. The following regressions were obtained (note: x1sq=x1*x1, x2sq=x2*x2, and x1x2=x1*x2).

Regression Analysis

```
The regression equation is
y = 58.0 + 0.104 x1 - 0.444 x2 - 0.0033 x1sq + 0.0257 x2sq + 0.0246 x1x2

Predictor        Coef       StDev          T        P
Constant      58.0406      0.1983     292.68    0.000
x1            0.10448     0.07790       1.34    0.213
x2            -0.4436      0.1725      -2.57    0.030
x1sq         -0.00330     0.01151      -0.29    0.781
x2sq          0.02566     0.04086       0.63    0.546
x1x2          0.02461     0.01668       1.48    0.174

S = 0.07460     R-Sq = 96.2%     R-Sq(adj) = 94.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         5     1.26584     0.25317     45.49    0.000
Residual Error     9     0.05008     0.00556
Total             14     1.31592
```

Regression Analysis

```
The regression equation is
y = 57.8 + 0.134 x1 - 0.267 x2

Predictor        Coef       StDev          T        P
Constant      57.8305      0.0644     898.44    0.000
x1            0.13389     0.01342       9.98    0.000
x2           -0.26707     0.02325     -11.49    0.000

S = 0.07351     R-Sq = 95.1%     R-Sq(adj) = 94.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     1.25107     0.62554    115.75    0.000
Residual Error    12     0.06485     0.00540
Total             14     1.31592
```

a. Test the hypothesis that the model without interactions is sufficient.
b. When testing the hypothesis that x1 is related to the response, do we need to include the x2sq term as well in our regression?
c. Which regression would you use to test the hypothesis that the coefficient associated with x1 is not zero. Why?
d. Test the hypothesis that the coefficient associated with x1 is not zero.

4. Suppose use data on 14 individuals to estimate the relationship between gender, height (inches) and weight (pounds). I've run two regressions, one with an interaction between height and gender (1 if male, 0 otherwise) and one without.

Regression Analysis

```
The regression equation is
wt = - 329 + 6.96 ht - 131 gender + 2.35 HtTimesGender

Predictor        Coef       StDev          T        P
Constant      -328.80       63.78      -5.16    0.000
ht             6.9625      0.9607       7.25    0.000
gender        -130.78       92.55      -1.41    0.188
HtTimesG        2.353       1.359       1.73    0.114

S = 9.186       R-Sq = 96.6%     R-Sq(adj) = 95.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     23795.0      7931.7     93.99    0.000
Residual Error    10       843.9        84.4
Total             13     24638.9
```

Regression Analysis

```
The regression equation is
wt = - 407 + 8.14 ht + 29.2 gender

Predictor        Coef       StDev          T        P
Constant      -406.79       49.10      -8.29    0.000
ht             8.1391      0.7385      11.02    0.000
gender         29.237       5.908       4.95    0.000

S = 9.987       R-Sq = 95.5%     R-Sq(adj) = 94.7%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2       23542       11771    118.03    0.000
Residual Error    11        1097         100
Total             13       24639
```

a. If your purpose is predicting a weight based on someone's height and gender, would you use the model with the interaction or the model without? Why?
b. Using the model with the interaction, draw the regression relationship between height and predicted weight for males and females on the same graph. (Suggestion: for some fixed heights, find predicted weight for each gender, plot and connect the dots appropriately.)
c. Does your plot in part b suggest the interaction, if it actually exists, is a strong one?
d. Explain, in plain English, what an interaction in this model would mean.

5. The following MINITAB output is based on n=25 observations on y=catch at intake (number of fish), x_1 =water temperature (degrees Celsius), x_2 =minimum tide height (m), x_3 =number of pumps running, x_4 = speed (knots), and x_5 =wind-range of direction (degrees) appeared in the article “Multiple Regression for Forecasting Critical Fish Influxes at Power Station Intakes” (*J. Applied Ecol.* (1983): 33-42).

Regression Analysis

```
The regression equation is
y = 101 - 2.84 x1 + 6.29 x2 - 23.6 x3 + 1.44 x4 + 0.0860 x5

Predictor        Coef       StDev          T        P
Constant       101.16       45.81       2.21    0.040
x1             -2.844       1.211      -2.35    0.030
x2              6.292       5.415       1.16    0.260
x3             -23.56       10.42      -2.26    0.036
x4             1.4387      0.6479       2.22    0.039
x5            0.08602     0.05811       1.48    0.155

S = 11.24       R-Sq = 34.0%     R-Sq(adj) = 16.6%

Analysis of Variance

Source             DF          SS          MS         F        P
Regression          5      1233.8       246.8      1.95    0.132
Residual Error     19      2400.2       126.3
Total              24      3634.0
```

a. Test the hypothesis that all of the betas are zero versus the alternative that at least one is nonzero.
b. Does it appear that water temperature is related to catch at intake? Explain.
c. Which variable would be eliminated first by backwards deletion?
d. It is interesting to note that both backwards deletion and forward selection suggest a model with **no** terms (i.e. the model without any explanatory variables). Why might this be the case?

Instructions:

- Write your name on your exam book. Write your name on your formula sheet. You will need to turn both in.
- Work independently. If you would like me to rephrase any questions, just ask.
- Answer all 4 questions.
- Each question is worth the same number of points. Sub-parts within each problem will each have about the same number of points.

1. Question 1.
   a. If it is believed that two factors (say, percentage of sand and grain size of the sand) affect the response (say, the strength of the concrete produced), explain why a 2^2 designed experiment is preferable to two separate, one-way experiments, one for each factor separately.
   b. Draw an "interactions plot" (displaying the means for each level of each factor) that shows no interaction between the two factors.
   c. Explain, in **plain English** what an interaction in a two way ANOVA means.
   d. Explain why a the main effects in a two-way ANOVA are more easily explained when there are no interactions. (I.e. explain why the main effects are not easy to interpret when there is an interaction.)
   e. In a 2^2 designed experiment, explain why an interaction cannot be tested unless there is more than one observation at each combination of treatment levels.

2. This question involves the following ANOVA table which was generated to analyze the results on an 2^3 experiment designed to determine whether factor C is related to the response. (Maybe it would be useful to view factors A and B as covariates which we already know are related to the response.)

source	df	ss	ms	f	p
a			157.960	179.704	0.0000
b		104.520	104.520	118.908	0.0000
c		3.780	3.780		
a*b		5.310	5.310		
a*c		0.110	0.110		
b*c		0.870	0.870		
a*b*c		0.410	0.410		
error	8	7.030			
total	15	279.990			

   a. Fill in the missing values in this table and indicate which of these F statistics are "statistically significant" at the 0.05 level. (Note: you should be sure to reproduce all 9 rows in this table in your exam book.)
   b. Should this ANOVA table be used to test the hypothesis that C is related to the response? If "yes", perform that test, if "no" explain why this table is inappropriate.
   c. Use an appropriate hypothesis test to determine whether at least one of the A*C, B*C or A*B*C interactions is statistically significant.
   d. Create a new ANOVA table by eliminating any interactions that are not statistically significant and use it to test the hypothesis that C is related to the response.

3. This question is about blocking.
   a. Explain why blocking may be useful in the design of an experiment.
   b. Describe a situation where a Latin Square design may be preferable to a simpler design where only one factor is used for blocking.
   c. Generate a 7×7 Latin Square for the 7 levels of treatment, {A, B, C, D, E, F, G}. Explain how this one Latin Square could be used to as a basis for generating a random allocation of these seven treatments across the 49 plots in the square.
   d. Explain how a Randomized Incomplete Block Design (RIBD) can be used if the size of the blocking unit does not allow for each treatment to have at least one replication within each block. (You may find it easier to explain how three treatments should be assigned when the blocks contain two plots each.)

4. An article in the June 1, 2000 copy of **the Tribune** reported on a recent analysis of 14,000 families in a welfare experiment in Minnesota. Half of the families were randomly assigned to a treatment where families were assisted via help finding jobs and a continuation of some benefits even after the participants had obtained employment. (That's the carrot part of this experimental carrot and stick welfare plan.) The other participating families received the traditional welfare that Minnesota provides. The article says that "among participants in the program, the number of marriages increased, existing marriages became more stable, domestic abuse dropped and children's behavior and academic performance improved." It appears that the only drawback of this program is that it cost an additional $2000 to $4000 more per family each year. Suppose you are in charge of setting up and analyzing the data from an experiment like this one that will measure the effect of this new sort of welfare system on the academic performance of the children involved. We are going to use the annual test scores of these children as our response variable.
   a. List one possible factor that we might want to use for blocking.
   b. Explain why randomization is important.
   c. List at least two covariates that are likely related to the response (test scores) that we should control for in our analysis.
   d. Which statistical method that we've discussed this term could be used to analyze such data appropriately? Explain briefly why your suggested method is appropriate.