36-202 STATISTICAL METHODS - Final
May 14, 2007

You must show your work and explain your steps in order to get full credit.

You should always comment on the numerical results.

You may use one **two-sided** sheet of notes (8.5 by 11 inches) and a calculator. You may not share a calculator, pencil, paper or anything else during the exam.

Your ANDREW ID: □□□□□□□□□

Your First Name: ____________________ **Your Last Name:** ____________________

Your Section: ____________________

Your Signature: ______________________________

Grader use:

Problem	Total	Correct
1		
2		
3		
4		
5		
6		
7		
8		
9		
Total		

***** Do not turn this page until instructed to do so.*****

PROBLEM 1:

In class we talked about several possible designs for examining the relationship between two categorical explanatory variables (factors), and a quantitative response:

- The two-factor between-subjects design ("regular" two-way ANOVA)
- The two-factor repeated measures design (two-way repeated measures ANOVA)
- The mixed between-within subjects design.

(a) Assume that a certain study can be done using either one of these three designs. Further assume that in this study:
- $Factor\ A$ has 2 levels
- $Factor\ B$ has 5 levels, and
- there are 12 subjects in each of the levels' combinations.

What is the total number of subjects that will be needed for the study using each one of the designs above? Explain.

Note: Show **both** possibilities for the mixed design (so you'll have 4 designs in total).

(b) Here is a specific example of a study that has two factors like those mentioned in (a):
A researcher is interested in studying how the variability in students' GPA can be explained by the student's `Gender` (M, F) and `Race` (which has 5 categories).

Which of the four designs which you described in part (a) *can* be used to conduct this study? Explain.

PROBLEM 2:

At the end of the semester, an instructor would like to analyze the data in her grade-book. In particular, the instructor would like to examine how much of the variability in the students' final exam scores can be explained using all the other variables in the grade-book as explanatory variables. The variables that are in the grade-book are:

HW1	score for homework assignment 1
HW2	score for homework assignment 2
HW3	score for homework assignment 3
HW4	score for homework assignment 4
HW.AVE	the average HW score of HW1-HW4
MID	score on the midterm
EX.CRED	whether the student turned in the extra credit assignment (1=yes, 0=no).
FINAL	final exam score

The instructor "throws" all the data into a statistical package, and asks it to conduct a multiple regression analysis using `FINAL` as the response, and *all* the other seven variables as explanatory. In return, the statistical package gives the following message: "The analysis could no be completed". Frustrated, the instructor comes to you, the statistical consultant for help.
Explain to the instructor the reason for the message she got, and suggest a "quick fix" to avoid the problem.

PROBLEM 3:

An experiment is conducted to compare the energy requirements of three physical activities: running, walking and bicycle riding. Eight subjects participated in the study, and each is asked to run, walk and bicycle a measured distance, and the number of kilocalories expended per kilometer is determined for each subject during each activity. Each subject completes the three activities in random order with time for recovery between activities. Each activity was monitored exactly once for each individual.

(a) The following two outputs are available. Circle the one which represents the appropriate way to analyze the results of this study as it was conducted. Explain briefly.

One-way ANOVA: Kilocalories versus Activity

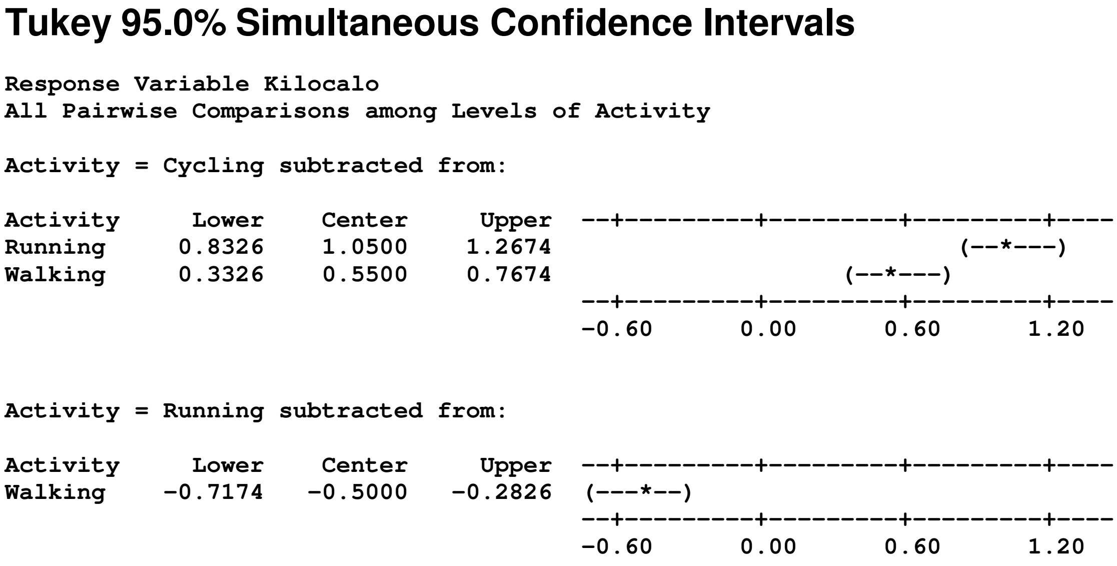
```
Analysis of Variance for Kilocalo

Source     DF        SS        MS        F        P
Activity    2    4.4133    2.2067    49.30    0.000
Error      21    0.9400    0.0448
Total      23    5.3533
```

ANOVA: Kilocalories versus Activity, Subject

```
Analysis of Variance for Kilocalo

Source      DF         SS         MS       F      P
Activity     2    4.41333    2.20667   79.90  0.000
Subject      7    0.55333    0.07905    2.86  0.045
Error       14    0.38667    0.02762
Total       23    5.35333
```

(b) State the hypotheses that are being tested in this study, and state your conclusions (in context) based on the output you chose in (a).

(c) The following pairwise comparisons output is available:



Tukey 95.0% Simultaneous Confidence Intervals

```
Response Variable Kilocalo
All Pairwise Comparisons among Levels of Activity

Activity = Cycling subtracted from:

Activity     Lower    Center     Upper  --+---------+---------+---------+----
Running     0.8326    1.0500    1.2674                            (--*---)
Walking     0.3326    0.5500    0.7674                    (--*---)
                                        --+---------+---------+---------+----
                                        -0.60      0.00      0.60      1.20


Activity = Running subtracted from:

Activity     Lower    Center     Upper  --+---------+---------+---------+----
Walking    -0.7174   -0.5000   -0.2826  (---*--)
                                        --+---------+---------+---------+----
                                        -0.60      0.00      0.60      1.20
```

Briefly summarize what the pairwise comparison analysis tells you about the *nature* of the activity effect on the number of kilocalories expended per kilometer by addressing the following:
(i) Which activities are significantly different in terms of energy expenditure?
(ii) Rank the activities from "least energy demanding" to "most energy demanding".

PROBLEM 4:

A study was done on the relationship between gender and piercing among high-school students. A sample of 1000 students was chosen, then classified according to gender, and whether or not they had any of their ears pierced. The results are summarized in the following 2×2 table:

	Piercing?		
Gender	Yes	No	Total
Female	576	64	640
Male	72	288	360
Total	648	352	1000

(a) Based on the observed data:

(i) What are the estimated odds that a female high-school student has pierced ears?

(ii) What are the estimated odds that a male high-school student has pierced ears?

(b) If we were to run binary logistic regression for this data with the response `Piercing (1=yes, 0=no)` and the explanatory `Gender (1 = female, 0 = male)`, the estimated logistic regression equation would be:

$$\hat{p} = \frac{1}{1 + e^{-(-1.3863 + \hat{\beta}_1 \cdot Gender)}}$$

Find the value of $\hat{\beta}_1$. (**Hint: use part (a)**)

PROBLEM 5:

Prior to the 2000 presidential elections, the National American Election Survey (NAES) asked the following question:

Do you favor or oppose the death penalty for persons convicted of murder?

1. *Favor Strongly*
2. *Favor not strongly*
3. *Oppose not strongly*
4. *Oppose strongly*

After the elections, each respondent was contacted again and asked who he/she voted for.

In this problem we will use data from the NAES to investigate whether a persons opinion on the death penalty is related to who the person voted for in the 2000 election. More specifically, we will examine whether Bush voters tend to favor the death penalty more than Gore voters.

We will therefore treat `Opinion on the death penalty` as the response variable (Y) which has a 4 point ordinal scale as given above, and the variable `Vote2000` as the explanatory (X) where 1=Bush, 0=Gore.

The following output are the results of fitting the ordinal logistic regression model to the data.

Ordinal Logistic Regression: Death.Penalty versus Vote2000

```
Response Information

Variable       Value  Count
Death.Penalty  1        296
               2         98
               3         68
               4         80
               Total    542

Logistic Regression Table

                                                   Odds     95% CI
Predictor        Coef   SE Coef       Z       P  Ratio  Lower  Upper
Const(1)    -0.303459  0.120599   -2.52   0.012
Const(2)     0.533013  0.122238    4.36   0.000
Const(3)      1.33891  0.139501    9.60   0.000
Vote2000     0.941989  0.168019    5.61   0.000   2.57   1.85   3.57


Log-Likelihood = -624.794
Test that all slopes are zero: G = 32.156, DF = 1, P-Value = 0.000

Goodness-of-Fit Tests

Method    Chi-Square  DF      P
Pearson      1.75495   2  0.416
Deviance     1.77014   2  0.413
```

(a) Write down the ordinal logistic regression *model* that is fitted to the data. Clearly define the probabilities that appear in the model.

(b) The results of fitting the model you specified in (a) to the data appear in the output on the previous page. Based on the output, do we have any reason to suspect that this model does not fit the data well? (Be sure to mention which part of the output you are using to answer this question).

(c) Use the output to estimate the probability that:

(i) a Gore voter favors (either strongly or not strongly) the death penalty.

(ii) a Bush voter strongly opposes the death penalty.
(Hint: The complement rule says $P(Y \geq j) = 1 - P(Y \leq j - 1)$**)**

(d) Is the explanatory variable Vote2000 significant?
Write down the appropriate hypotheses, give the p-value and state your conclusions in context.

(e) To quantify the Vote2000 effect, interpret the value of the odds ratio $e^{\hat{\beta}_1}$ in context.

(f) Complete the sentence:
We are 95% confident that e^{β_1} falls between ______________ and ______________.

PROBLEM 6:

We've recently looked at the following example:

The director of admissions of a certain liberal arts college (College C) noticed that in recent years there has been an increase in the number of students who are accepted to the college but end up going to a different school. A more thorough investigation revealed that a lot of those student end up going to one of two other liberal arts colleges (College A and College B). The director of admissions would like to learn more about what type of students end up choosing one of the other colleges over College C, and gathers information about students who in the past five years have been admitted to all three colleges. For each student, the following variables (among others) were recorded:

- **HS.GPA** - the student's high-school GPA (out of 100)
- **Gender** - 1–female, 0–male.
- **Minority** - Whether the student is a minority student (1–yes, 0–no)
- **School** - Which college the student ended up choosing (A, B, or C).

The Nominal Logistic Regression model was fitted to the data, and the output is given below. Answer the question on the next page.

Nominal Logistic Regression: School versus Minority, Gender, HS.GPA

```
Response Information

Variable  Value  Count
School    C        150  (Reference Event)
          B        175
          A        125
          Total    450


Logistic Regression Table

                                                      Odds     95% CI
Predictor            Coef    SE Coef      Z      P   Ratio  Lower  Upper
Logit 1: (B/C)
Constant         -36.4825    4.11746  -8.86  0.000
Minority         0.492470   0.340194   1.45  0.148    1.64   0.84   3.19
Gender          0.0265998   0.258884   0.10  0.918    1.03   0.62   1.71
HS.GPA           0.406989  0.0456080   8.92  0.000    1.50   1.37   1.64

Logit 2: (A/C)
Constant         -1.59302    2.83483  -0.56  0.574
Minority          1.18380   0.294955   4.01  0.000    3.27   1.83   5.82
Gender          0.0819408   0.252307   0.32  0.745    1.09   0.66   1.78
HS.GPA          0.0122077  0.0322219   0.38  0.705    1.01   0.95   1.08


Log-Likelihood = -406.979
Test that all slopes are zero: G = 166.420, DF = 6, P-Value = 0.000


Goodness-of-Fit Tests

Method    Chi-Square   DF       P
Pearson      279.390  142  0.257
Deviance     181.564  142  0.464
```

(a) Briefly explain why the nominal logistic regression model is appropriate here (as oppose to the binary logistic regression model or the ordinal logistic regression model).

(b) Based on the output we can conclude that holding all other variables fixed, (circle one)

(i) for every one unit increase in `HS.GPA`, the probability that a student prefers college B over the other two colleges increases by 50%.

(ii) for every one unit increase in `HS.GPA`, the odds that a student prefers college B over the other two colleges increase by 50%.

(iii) for every one unit increase in `HS.GPA`, the probability that a student prefers college B over college C increases by 50%.

(iv) for every one unit increase in `HS.GPA`, the odds that a student prefers college B over college C increase by 50%.

(c) Based on the output we can conclude that holding all other variables fixed, (circle one)

(i) the odds that a non-minority student prefers college A over the other two colleges is 3.27 times the odds that a minority student prefers college A over the other two.

(ii) the odds that a minority student prefers college A over college C is 3.27 times larger than the odds that a non-minority student prefers college A over college C.

(iii) the probability that a minority student prefers college A over college C is 3.27 times larger than the probability that a non-minority student prefers college A over college C.

(iv) the odds that a non-minority student prefers college A over college C is 3.27 times larger than the odds that a minority student prefers college A over college C.

PROBLEM 7:

A study was done in order to be able predict p, the probability that a randomly selected respondent supports current laws legalizing abortion, using the respondent's age, gender (male, female), religious affiliation (Protestant, Catholic, or Jewish) and political party affiliation (Democrat, Republican, or Independent). The study reported the following estimated binary logistic regression equation:

$$\ln\left(\frac{\hat{p}}{1-\hat{p}}\right) = .11 - .10A + .16G - .57R_1 - .66R_2 + .47P_1 - 1.67P_2$$

where:

- A – age (in years)
- $G = 1$ if female and 0 if male
- $R_1 = 1$ if Protestant and 0 otherwise
- $R_2 = 1$ if Catholic and 0 otherwise
- $P_1 = 1$ if Democrat and 0 otherwise
- $P_2 = 1$ if Republican and 0 otherwise

(a) Find $e^{\hat{\beta}_1}$ (where $\hat{\beta}_1$ is the coefficient of A), and interpret this value in the context.

(b) Find $e^{\hat{\beta}_4}$ (where $\hat{\beta}_4$ is the coefficient of R_2), and interpret this value in the context.

(c) The first individual in our data is a 60 year old Jewish male who identifies himself as a Democrat. Use the estimated logistic regression equation to estimate the probability that this person supports the current laws legalizing abortion.
Comment: It might be useful to consider another form of the estimated logistic regression equation which would be more convenient.

(d) The second individual in our data is a 35 year old Catholic woman who identifies herself as Independent. Use the estimated logistic regression equation to estimate the probability that this person supports the current laws legalizing abortion.

(e) Assume now that the men in part (c) responded that he supports the current law legalizing abortion ($Y = 1$), that the woman in (d) responded that she opposes the current law ($Y = 0$). Classify this pair as either concordant or discordant and explain your answer.

PROBLEM 8:

A leisure researcher was interested in determining whether age and gender had any bearing upon the amount of time adults engaged in leisure activities.
- `Age` was categorized into 3 levels (1=Young adults, 2=middle-aged adults, 3=older adults),
- `Gender` was coded using M and F, and
- the response, `Time`, was measured in hours per week.

The researcher obtained a random sample of 10 individuals for each of the six explanatory variables' levels combinations, and plans on using two-way ANOVA to analyze the data.

The output is on pages 16-18 of this exam

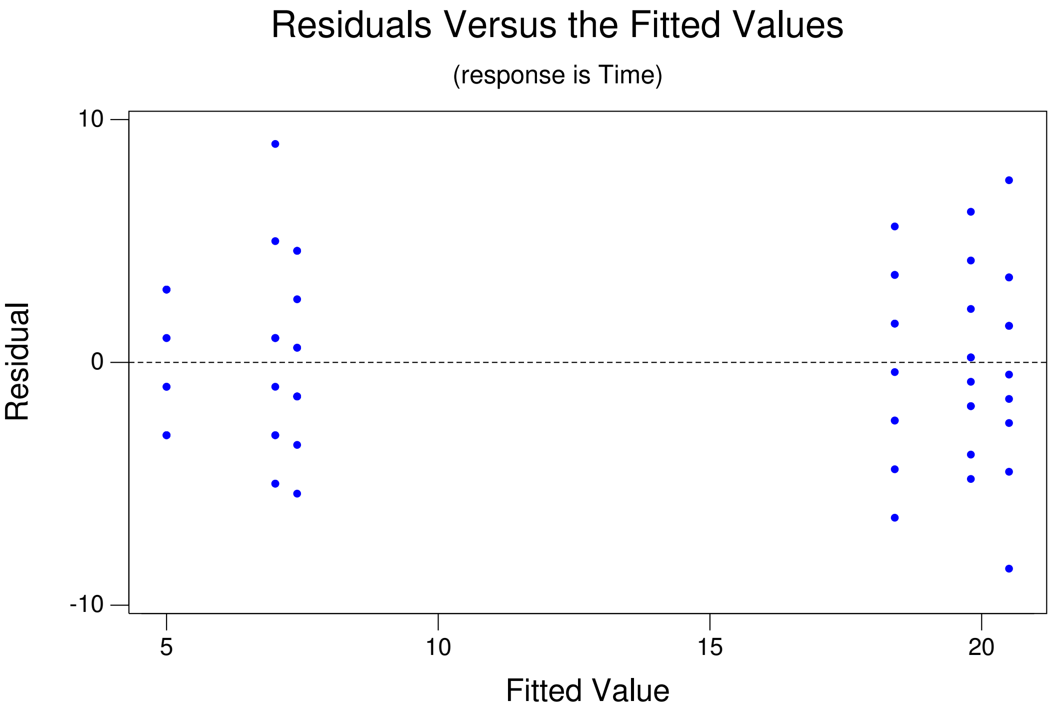
Answer the following questions:

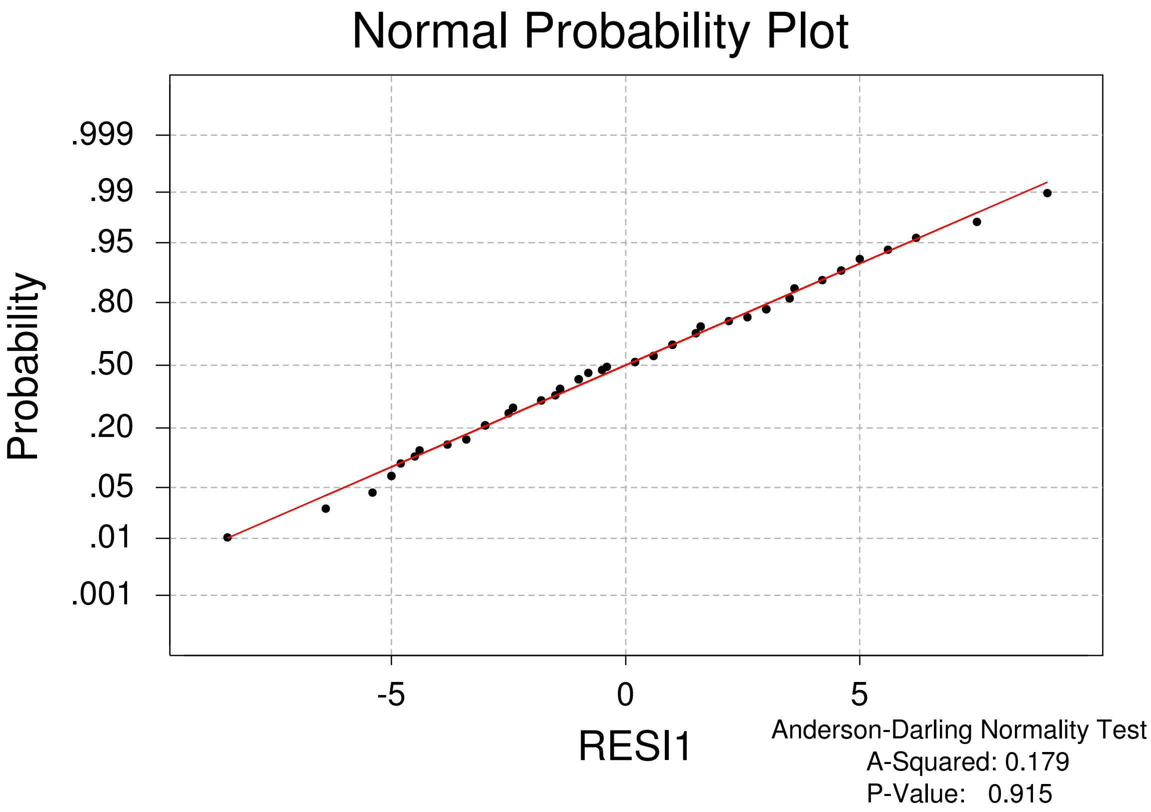
(a) Before carrying out the two-way ANOVA, check whether there are any violations of the model assumptions of normality and equal spread within the groups.
- Be sure to mention what part of the output you are using to check each assumption.
- If you find that an assumption is violated, be clear about measures you are going to take.

(b) Based on the ANOVA table, report on the significance/non-significance of the results. Support your answer by the appropriate p-value(s).

(c) Summarize the findings of this study.
- Be very clear about which plot(s) you are basing your summary on (Note, there are a few plots available, but not all are relevant or appropriate to use....)

(d) How much of the variation in `Time` can be explained by our model, altogether? (In other words, find R^2).


Residuals Versus the Fitted Values
(response is Time)
Residual
10
0
-10
5
10
15
20
Fitted Value

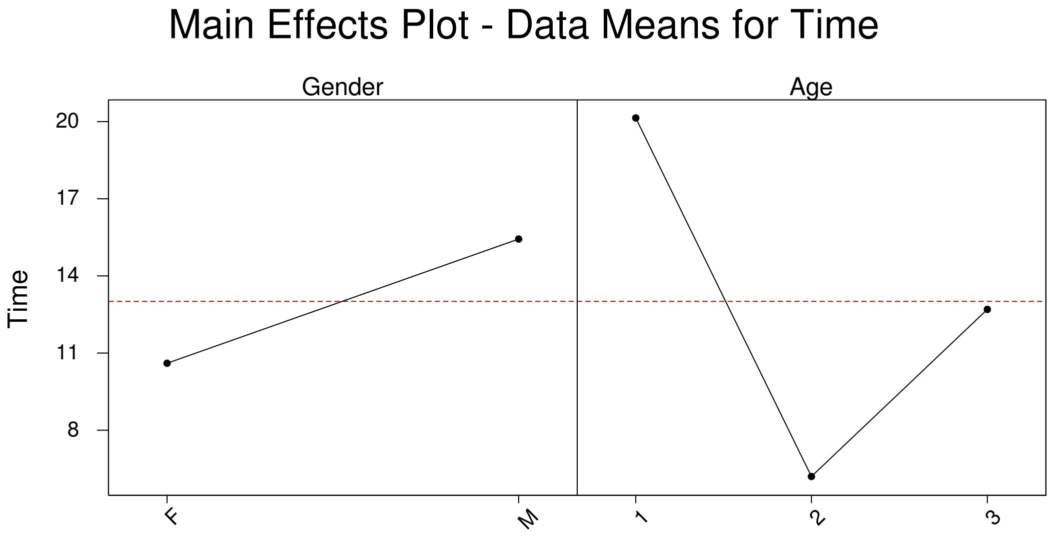

Normal Probability Plot
Probability
.999
.99
.95
.80
.50
.20
.05
.01
.001
-5
0
5
RESI1
Anderson-Darling Normality Test
A-Squared: 0.179
P-Value: 0.915

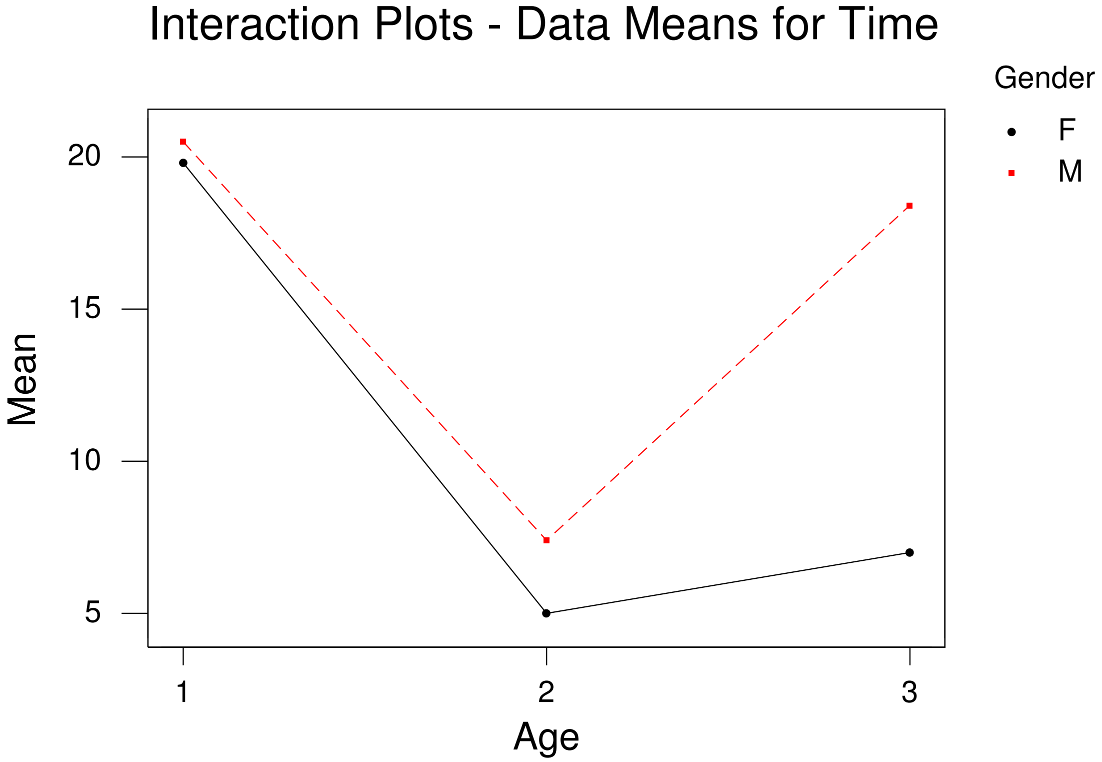
Tabulated Statistics: Gender, Age

```
Rows: Gender     Columns: Age

           1         2         3       All

 F    19.800     5.000     7.000    10.600
       3.425     2.539     4.447     7.500

 M    20.500     7.400    18.400    15.433
       4.552     3.273     3.864     6.966

 All  20.150     6.200    12.700    13.017
       3.937     3.105     7.116     7.579

  Cell Contents --
             Time:Mean
                  StDev
```

Two-way ANOVA: Time versus Gender, Age

```
Analysis of Variance for Time
Source        DF        SS        MS        F        P
Gender         1     350.4     350.4    24.93    0.000
Age            2    1949.0     974.5    69.34    0.000
Interaction    2     330.6     165.3    11.76    0.000
Error         54     758.9      14.1
Total         59    3389.0
```


Main Effects Plot - Data Means for Time
Gender
Age
Time
20
17
14
11
8
F
M
1
2
3


Interaction Plots - Data Means for Time
Gender
● F
■ M
Mean
20
15
10
5
1
2
3
Age

PROBLEM 9:

An interviewer asked a random sample of 45 students at a state university to decide whether each of the following acts should be considered a crime: aggravated assault, armed robbery, arson, atheism, auto theft, burglary, civil disobedience, communism, drug addiction, embezzlement, forcible rape, gambling, homosexuality, land fraud, lying, Nazism, payola, price fixing, prostitution, sexual abuse of children, sexual discrimination, shoplifting, strip mining, treason, vandalism. For each student, the interviewer determined the number of acts considered a crime (the response) and other information concerning the interviewee: years of college education, age, income of parents (in thousands of dollars), and gender(1=female, 0=male). These are the four explanatory variables.

The output is on pages 22-25 of this exam

Answer the following questions:

(a) If you had to choose only *one* of the four explanatory variable as a predictor for the response `Crime` (the number of acts which are considered a crime), which variable would it be? Explain your reasoning.

(b) What is the least squares regression equation that relates the variable you chose in (a) to the response? Interpret the value of b_1 ($\hat{\beta}_1$) in context.

We are now ready to fit the *multiple* linear regression model to the data (i.e., include all four explanatory variables simultaneously).

(c) Write down the multiple regression *model* that we are fitting to the data (including the assumptions about the deviations).

(d) Write down the least squares regression equation that relates `Crime` to the four explanatory variables, and interpret the value of b_4 (the coefficient of `Gender`) in the context of the problem.

(e) For each of the explanatory variables, state whether it is significant or not (in the presence of the other explanatory variables) and support you answer by the appropriate p-value.

(f) There are two *different* ways in which the multiple regression output indicates that a multicollinearity problem exists. What are they?

(g) Because of the multicollinearity problem, it is clearly not a good idea to keep the full model. Suggest a "reduced" model that you think is better, explain your reasoning, and write down the least squares regression equation for that model.

Regression Analysis: Crime versus Age

```
The regression equation is
Crime = 7.47 + 0.290 Age

Predictor        Coef     SE Coef          T        P
Constant        7.474       4.284       1.74    0.088
Age            0.2897      0.1765       1.64    0.108

S = 5.866       R-Sq = 5.9%      R-Sq(adj) = 3.7%
```

Regression Analysis: Crime versus College

```
The regression equation is
Crime = 13.5 + 0.29 College

Predictor        Coef     SE Coef          T        P
Constant       13.509       3.196       4.23    0.000
College         0.286       1.037       0.28    0.784

S = 6.042       R-Sq = 0.2%      R-Sq(adj) = 0.0%
```

Regression Analysis: Crime versus Income

```
The regression equation is
Crime = - 0.20 + 0.302 Income

Predictor        Coef     SE Coef          T        P
Constant       -0.196       1.661      -0.12    0.906
Income        0.30177     0.03270       9.23    0.000

S = 3.502       R-Sq = 66.5%     R-Sq(adj) = 65.7%
```

Regression Analysis: Crime versus Gender

```
The regression equation is
Crime = 12.0 + 5.98 Gender

Predictor        Coef     SE Coef          T        P
Constant       11.963       1.011      11.84    0.000
Gender          5.981       1.598       3.74    0.001

S = 5.252       R-Sq = 24.6%     R-Sq(adj) = 22.8%
```

Regression Analysis: Crime versus Age, College, Income, Gender

```
The regression equation is
Crime = - 10.8 + 0.432 Age - 0.02 College + 0.290 Income + 2.45 Gender

Predictor        Coef     SE Coef          T        P       VIF
Constant      -10.823       2.392      -4.52    0.000
Age            0.4324      0.2024       2.14    0.039       6.7
College        -0.024       1.221      -0.02    0.984       7.5
Income        0.29025     0.03142       9.24    0.000       1.7
Gender         2.4542      0.8747       2.81    0.008       1.2

S = 2.601       R-Sq = 82.8%     R-Sq(adj) = 81.1%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         4     1301.62      325.41     48.09    0.000
Residual Error    40      270.69        6.77
Total             44     1572.31

Source       DF      Seq SS
Age           1       92.68
College       1      263.39
Income        1      892.28
Gender        1       53.28
```

Best Subsets Regression: Crime versus Age, College, Income, Gender

```
Response is Crime

                                                  C
                                                  o I G
                                                  l n e
                                                  l c n
                                                A e o d
                                                g g m e
Vars   R-Sq    R-Sq(adj)          C-p         S e e e r

   1   66.5         65.7         36.9    3.5023       X
   1   24.6         22.8        134.2    5.2516         X
   2   79.3         78.3          9.1    2.7843   X   X
   2   78.1         77.1         11.8    2.8609     X X
   3   82.8         81.5          3.0    2.5695   X   X X
   3   80.8         79.4          7.6    2.7121     X X X
   4   82.8         81.1          5.0    2.6014   X X X X
```

Regression Analysis: Crime versus Age, Income

```
The regression equation is
Crime = - 11.3 + 0.432 Age + 0.320 Income

Predictor        Coef     SE Coef          T        P       VIF
Constant      -11.338       2.552      -4.44    0.000
Age           0.43164     0.08459       5.10    0.000       1.0
Income        0.32019     0.02624      12.20    0.000       1.0

S = 2.784       R-Sq = 79.3%     R-Sq(adj) = 78.3%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         2     1246.71      623.36     80.41    0.000
Residual Error    42      325.60        7.75
Total             44     1572.31

Source       DF      Seq SS
Age           1       92.68
Income        1     1154.04
```

Regression Analysis: Crime versus Age, Income, Gender

```
The regression equation is
Crime = - 10.8 + 0.429 Age + 0.291 Income + 2.45 Gender

Predictor        Coef     SE Coef          T        P       VIF
Constant      -10.822       2.362      -4.58    0.000
Age           0.42872     0.07807       5.49    0.000       1.0
Income        0.29058     0.02630      11.05    0.000       1.2
Gender         2.4511      0.8500       2.88    0.006       1.2

S = 2.569       R-Sq = 82.8%     R-Sq(adj) = 81.5%

Analysis of Variance

Source            DF          SS          MS         F        P
Regression         3     1301.62      433.87     65.72    0.000
Residual Error    41      270.69        6.60
Total             44     1572.31

Source       DF      Seq SS
Age           1       92.68
Income        1     1154.04
Gender        1       54.90
```