Biostatistics 201A, Midterm Practice Problems

General Comments:

- Since I have only taught this class once before, I do not have a lot of complete exams exactly tailored to the material we have covered so far. I have posted last year's midterm and also included in this file a mix of problems from other courses that I think will be useful. I do not guarantee that I have here a problem on every subject that is fair game for the exam or that the balance is exactly right. In particular this set is currently a bit short on t-tests and power. Let me know if you want me to post more problems in this area.
- In addition to (in fact even more important than) the practice problems it is worth reviewing the problems from HW 1-4 as the assignments most accurately reflect what I think are the key topics we have covered so far.
- You can easily create more practice problems for yourself by taking any of these printouts and interpreting all the parts of them. I have by no means asked every possible question about each printout!
- The exam will be closed book. However, you may use two pages, front and back, of notes and formulas. Write your answers on the exam sheets. If you need more space, continue your answer on the back of the pages. Normal, t and F distribution tables will be provided for you at the end of the exam and I will have scratch paper available if you want it.
- **You must show your work** on the exam to obtain full credit. If you use a result from class, state what result you are using. If you can't complete a problem for any reason, explain what concepts are at issue, and how you would attack the problem. If you can't work out a number you need for a later part of a problem give it a symbol and show how you would do the calculations with a symbol in place of the missing number. It is, in any case, a good idea to explain **briefly** what your reasoning is in English. If I can't tell that you understood what you were doing, I can't give you credit, particularly if you get the wrong numerical answer. HAPPY STUDYING!

Inference, Including Estimzation, CIs, Tests and Power

(1) Charity Donations: A charity suspects one of its volunteers of stealing donations. The volunteer is going door to door asking for contributions. Past experience shows that honest volunteers collect on average \$3 per household. However, obviously, the actual amount varies for each household. In the last week the volunteer visited 500 households and the end of the week reported his total donations as \$1149 with a standard deviation of the amount collected from each person of $s = 5$.

(a) What is your best estimate of the volunteers average donation per household during this period?

(b) Find a 95% confidence interval for this volunteer's average donation based on the data from the past week. Based on this interval does the volunteer appear to be typical? Discuss.

(c) Conduct the hypothesis test an investigator would use to prove the volunteer is stealing. Be careful to state the null and alternative hypotheses, both mathematically and in English, with a justification and explain your final conclusion. Use $\alpha = .025$. and compare your result to that of part (b).

(d) Why did I say to use $\alpha = .025$ in part (c) when you computed a 95% CI in part (b)?

(e) Do the results of parts (b) and (c) prove the volunteer is stealing? Give two possible alternative explanations.

(f) What is the effect size for the difference between the volunteer being studied and the agencies typical volunteers? How big would you consider this effect to be?

(g) What is the minimum detectable effect size for this study for a two-sided test with $\alpha = .05$ and 80% power? (Note–you need to use STATA to do this as stated. On an exam I would have to give you a printout to interpret.) In general do you consider the power here to be high? Discuss.

(h) Suppose instead of comparing the volunteer from the previous parts of this problem to previous population values we instead had two volunteers, one of whom was the the one above and the second a colunteer who had averaged $\bar{X} = \$3$ per donation with s = 4 for 25 donations. (i) Find a confidence interval for the difference in means between the two volunteers. (ii) Perform an appripriate hypothesis test to see if the amounts they collect are different (iii) Give an estimate of the effect size for the difference in their average collections.

Classical ANOVA

(1) Honey I Shrunk the Tumor: Dr. Clever is studying a rare form of cancer. People who have this form of cancer have traditionally received either standard chemotherapy (C) or aggressive chemotherapy (AC). Dr. Clever has himself developed a new medication (I) that is injected directly into the tumor site. Dr. Clever is interested in knowing how much the different treatments shrink people's cancer tumors. He has collected data on tumor size before and after treatment for 13 subjects who have received standard chemotherapy, 25 subjects who have received agressive chemotherapy and 25 subjects who have received his new injectable treatment. For each person he has calculated the percentage reduction in tumor size, Y. For instance, Y = 50 corresponds to a 50% reduction in tumor size. An ANOVA printout for his data is shown below.

```
. oneway PercentShrunk Treatment, tabulate

           |     Summary of PercentShrunk
 Treatment |     Mean   Std. Dev.    Freq.
-----------+------------------------------------
         C |      37       20         13
        AC |      38       20         25
         I |      50       10         25
-----------+------------------------------------
     Total |    42.56    17.55        63


Number of obs =     63     R-squared     =  0.121
Root MSE      =  16.73     Adj R-squared =  0.091

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups       2305.56        2   1152.78        4.12      0.021
 Within groups      16800.00       60    280.00
------------------------------------------------------------------------
    Total           19105.56       62    308.15
```

(a) Based on this data is there evidence the different treatments shrink the tumors by different average amounts? Justify your answer by performing an appropriate **overall** hypothesis test. State the mathematical hypotheses in the classical ANOVA framework, give the p-value, and your real-world conclusions.

The test statistics and p-values for pairwise comparisons of the group means in this ANOVA are given below. Use them to answer parts (b)-(d).

Pair	t_{obs}	p-value
C vs AC	.175	0.862
C vs I		
AC vs I	2.536	0.0138

(b) One of the rows is blank. Write down the mathematical null and alternative hypothesis corresponding to the missing test, compute the test statistic and give as close an approximation as you can to the p-value.

(c) Based on the ANOVA printout, the table and your answer to part (b), order the three treatments from most to least effective (in terms of percentage shrinkage of the tumor), clearly indicating which differences are significant and which are not at $\alpha = .05$.

(d) Suppose you used the Bonferroni correction to adjust for the multiple comparisons in part (c). Say what the new significance level, α^*, for the individual tests would be and indicate how your

answer to (c) would change (if at all).

(e) Dr. Clever is interested in proving that his new injection therapy works **better** than either of the chemotherapy treatments.

(i) Write down an appropriate linear combination, LC, for the comparison he wishes to do. (You may assume that the two chemotherapy treatments are equally common in the population of interest.)

(ii) Give your best estimate of LC and the corresponding standard error.

(iii) Use these numbers to compute a **90% confidence interval** for the linear combination and give a brief interpretation of it. Your interpretation should incorporate the numerical values of both ends of the interval and also address the question of interest to Dr. Clever.

(f) Optional Bonus Dr. Clever's graduate student, Denny Dull, suggests that instead of looking at the percentage shrinkage of the tumors Dr. Clever should do an ANOVA with 6 groups giving the mean tumor sizes before and after treatment for each of the 3 treatment groups. (i) Explain what regression assumption would be violated if Dr. Clever performed the analysis this way and (ii) give one other potential problem with this approach.

(g) Optional Bonus) Suppose, continuing part (e) that Dr. Clever wants to prove his injection therapy shrinks the cancer tumors by **at least 5% more** than the average of the chemotherapy treatments. Write down the hypotheses for this test mathematically and in words, compute the test statistic, and give an approximate p-value and real-world conclusions using $\alpha = .05$.

(2) Fun with Genetic Testing: You are studying whether a group of genes is associated with an elevated risk of breast cancer. For each person you record X, and indicator of whether or not the person got breast cancer and Y_j, the expression level (a continuous measure) associated with gene j for a large number of different genes.

(a) Explain what methods we have learned could be used to tell with expression level of a given gene is associated with breast cancer.

(b) Imagine that you wanted to identify the subset of genes tested that were actually related to cancer. Suppose you were testing 10 genes (so $j = 1, 2, \ldots, 10$), what approaches could you use to ensure the overall accuracy of your answers?

(c) Now imagine that instead of the 10 genes in part (a) you were testing 10,000 genes. Now what would be a reasonable approach to ensuring overalla ccuracy? Discuss.

Simple Linear Regression

(1) Mostly Mozart: Dr. Smart believes that the mother's drinking during pregancy will have a long term negative effect on child's mental development while listening to classical music will have a positive effect. She has therefore conducted a study in which she followed 102 pregnant woman and recorded both X_1, the average number of drinks they had each day during the pregnancy and X_2, the number of minutes they listened to classical music each day. She then went back when the children were 7 years old and recorded their scores, Y, on a standard IQ test. (For reference, an average IQ score is around 100 while scores below 70 correspond to mental retardation and scores around 160 are thought to represent genius.) Dr. Smart planned to perform two simple linear regressions, one of IQ on mother's alcohol consumption and one of IQ on mother's music listening, along with corresponding correlation and covariance calculations. The STATA printouts for the music analysis are given below. However those for the alcohol analysis seem to have been lost and all that is available are some summary statistics. Use this information to answer the questions on the following pages.

Overall: $n = 102$, $\bar{Y} = 95$

For the Alcohol Analysis: $\bar{X}_1 = 1$, $SCP = -2000$, $SSX = 400$, $SST = SSY = 20000$

For the Music Analysis:

Correlation:

```
. corr IQ Music
(obs = 1-2)
             |       IQ    Music
-------------+------------------
          IQ |   1.0000
       Music |   0.3000   1.0000
```

Covariance

```
.corr IQ Music, c
             |       IQ    Music
-------------+------------------
          IQ |   198.02
       Music |   356.40    7128.7
```

Regression:

```
. reg IQ Music

      Source |       SS       df       MS              Number of obs =     102
-------------+------------------------------           F(  1,   100) =    9.89
       Model |     1800        1     1800              Prob > F      =   0.002
    Residual |    18200      100      182              R-squared     =   0.090
-------------+------------------------------           Adj R-squared =   0.081
       Total |    20000      101   198.02              Root MSE      =   13.49

------------------------------------------------------------------------------
          IQ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Music |      0.05     0.016      3.145   0.001        0.0185     0.0815
       _cons |     93.50     1.418     65.924   0.000       90.6861    96.3139
------------------------------------------------------------------------------
```

(a) Which is stronger, the relationship between IQ score and maternal alcohol consumption or the relationsip between IQ score and maternal music listening? Explain your reasoning and show any necessary calculations.

(b) Find b_0 and b_1, the estimates for the intercept and slope, of the simple linear regression of IQ (Y) on maternal alcohol consumption (X_1) based on these data.

(c) Give the units and real-world interpretations of b_0 and b_1 for the regression of IQ (Y) on maternal music listening (X_2) and say briefly whether they make real-world sense. Your answer should incorporate the actual numerical values of the coefficients.

(d) Silly Sally, a graduate student at the University of the Statistically Challenged, who is pregnant with her first child, decides on the basis of this study that she will have classical music playing in her house 24 hours a day, 7 days a week. According to the IQ-music regression model what is the predicted IQ for her child? Do you think this prediction is reliable? Explain.

(e) What is Sally assuming when she makes the decision to play non-stop classical music? Is her assumption justified? If so, explain why and if not give an appropriate example (in the context of this problem!) to back up your argument.

(f) Suppose instead of fitting two simple linear regressions we fit a single model that used both alcohol consumption and music listening to predict IQ. Relative to their values in the simple linear regression, indicate (by circling your choice) whether you would expect SST, SSR and SSE in the new model to be larger, smaller or stay the same. Briefly explain your reasoning.

SST	Increase	Decrease	Stay the Same
SSR	Increase	Decrease	Stay the Same
SSE	Increase	Decrease	Stay the Same

(2) Fast Stats on Fast Food: Dr. Nutts has selected n=62 children from urban neighborhoods in the city of Los Seraphim. For each child she has recorded Y, the average amount the child eats each day in **hundreds** of calories, and X, the number of fast food restaurants within 3 miles of the child's house. Some data and a simple linear regression printout from her study are given below, although a few values seem to be missing. If you need a missing value and can't figure out how to compute it, simply explain how you would use it to answer the question if you had it.

$\bar{Y} = 21.2$ $\bar{X} = 5$ $SSX = 202$ $n = 62$

```
. regress Calories Restaurants

      Source |       SS       df       MS            Number of obs   =      62
```

```
-------------+------------------------------        F(  1,    60)   =   37.10
       Model |     70.2       1      70.2           Prob > F      = 0.00000004
    Residual |    113.6      60       1.9           R-squared       =   ?
-------------+------------------------------        Adj R-squared   =   ?
       Total |    183.8      61       3.0           Root MSE        =   1.376

------------------------------------------------------------------------------
    Calories |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 Restaurants |      .590       .097      6.09  0.00000004       .396     .783
       _cons |   18.251       .518     35.26  0.00000000     17.215   19.286
------------------------------------------------------------------------------
```

(a) Is there a significant **positive** linear relationship between the number of fast food restaurants in a child's neighborhood and the number of calories they consume? Give the the p-value and your real world conclusions for an appropriate test using $\alpha = .05$. (You do NOT need to write out all the details of the test.)

(b) Does number of fast food retaurants explain a high **percentage** of the variability in number of calories consumed? Explain briefly, showing any necessary calculations. (You do NOT need to use more than one number to justify your interpretation.)

(c) Does the number of fast food retaurants in their neighborhood do a good job of **predicting** the number of calories a child consumes? Carefully justify your answer.

(d) Find an interval which you can be 95% sure contains the average calorie consumption of children in a neighborhood with 10 fast food restaurants. Show your work.

(e) Standard guidelines suggest that very active children aged 9-13 need approximately 2200 calories per day, while less active children need fewer calories. Based on this and your answer to part (d) does it seem that children in neighboorhoods with 10 fast food restaurants have unhealthy diets on average? Explain briefly.

(f) Optional Bonus Suppose you had measured Y in calories and X in dozens of restaurants. What would the resulting regression equation have been? Show your work.

(3) Television Ads: You own a chain of stores that sells television sets and you want to know whether your advertising is increasing your sales. Let Y be the number of TVs you sell in a given month, and let X be the amount of money you spend on advertising in a given month in thousands of dollars. You have data on advertising expenditures and sales for n=42 months and have fit a simple linear regression of Y on X. The printout for this regression is given below along with a few useful summary statistics. Use it to answer the following questions:

```
The regression equation is
TV Sales = 48.4 + 10.2 Ad-Spending
```

```
Parameter Estimates

Predictor       Coef       Stdev    t-ratio        p
Constant       48.40       17.61       2.75    0.009
Ad Spending  10.2457      0.5224      19.61    0.000

Root MSE = 38.54       R-sq = 90.6%     R-sq(adj) = 90.3%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      571411      571411    384.70    0.000
Error        40       59413        1485
Total        41      630824
********************
Summary Statistics For Number of Televisions Sold Per Month

                N     MEAN   MEDIAN    STDEV      MIN      MAX
TV Sales       42    373.5    363.0    124.0    146.0    609.0
```

(a) What **percentage of variability** in television sales is explained by advertising expenditures? Does the model do a good job in this respect? Explain.

(b) Do advertising expenditures give good **predictions** for the number of television sales? Briefly justify your answer using appropriate numbers from the printouts.

(c) Is there a **significant linear relationship** between sales, Y, and advertising, X? Justify your answer by performing either a T test or an F test (your choice), making sure to give the null and alternative hypotheses both mathematically and in words. Also give the test statistic, the p-value, say whether or not you reject the null hypothesis and why, and state your real world conclusions. (Use $\alpha = .005$)

(d) Find a 99% confidence interval for β_1, the slope of the regression line, and briefly explain what it tells you about the relationship between advertising and television sales.

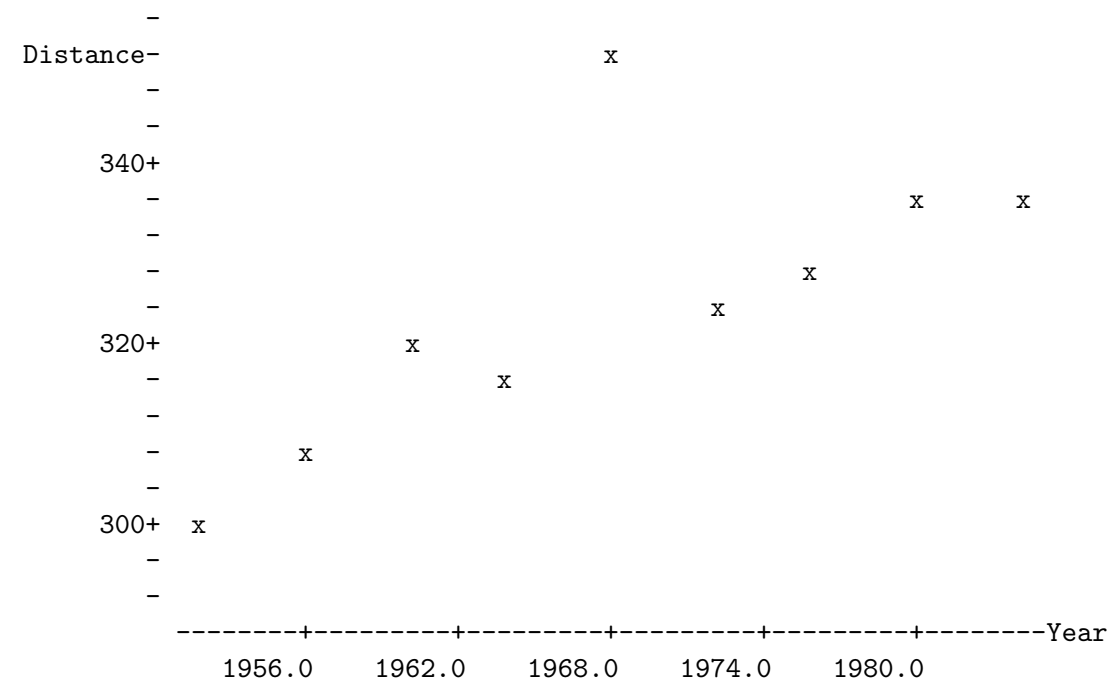
(e Suppose your company makes a \$100 profit per television sold BEFORE taking advertising costs into account. According to your **best estimate**, do the ads appear to be paying for themselves? Can you be 99% (really 99.5%) sure? Explain briefly.

(4) Leaping Into the Future

In the modern Olympic era, performances in track and field have been steadily improving. The table below gives the winning distance (in inches) for the Olympic long jump from 1952 to 1984. Below is a regression printout for a simple regression of distance on year. Use the printout to answer the following questions.

Year	Distance
1952	298
1956	308.25
1960	319.75
1964	317.75
1968	350.5
1972	324.5
1976	328.5
1980	336.25
1984	336.25

Scatterplot



Regression Analysis

```
. reg Distance Year

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =    9.21
       Model |  1137.52604     1  1137.52604           Prob > F      =  0.0190
    Residual |  864.973958     7  123.567708           R-squared     =  0.5681
-------------+------------------------------           Adj R-squared =  0.5063
       Total |      2002.5     8    250.3125           Root MSE      =  11.116

------------------------------------------------------------------------------
    Distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Year |   1.088542   .3587706     3.03   0.019     .2401839    1.936899
       _cons |  -1817.833   706.0703    -2.57   0.037    -3487.424   -148.2423
------------------------------------------------------------------------------
```

(a) Give the units and interpretation of b_1 in the simple regression model.

(b) What proportion of the variability in distance is explained by year using the simple linear regression model? Does the model do a good job in this respect?

(c) Does the simple linear regression model do a good job of predicting the Y values? Make sure you justify your answer.

(d) Is there a significant linear relationship between years and distance? Justify your answer using an appropriate test.

(e) In 1968, the Olympics were held in Mexico City, and many records were set, probably due to the high altitude. A point like this is called an outlier. Explain what would happen to your answers to (b)-(d) if this point were removed.

(5) Computer Chaos:

You have been hired as a statistical consultant by a large hardware store. They are interested in knowing how their sales of fans depend on the weather. They have presented you with data from the previous summer. Their data consists of two variables, Y, the number of fans sold in each week, and X the hottest temperature during that week. They have given you data for n=12 weeks. During those weeks the average temperature was found to be $\bar{X} = 80$ and the average number of fans sold per week was $\bar{Y} = 160$. You have further managed to calculate from your data that $SCP = 200, SSX = 100$, and $RMSE = 4$. You have gotten sick of doing the calculations by

hand and decided to use a computer. Unfortunately (what a shock) the program is malfunctioning and your printout has a lot of blanks. In this problem you will fill in the blanks and answer some questions for the hardware store. **Note:** It is possible to completely answer parts (b)-(g) even if you can't fill in a single number in the printout, so don't give up on them!!

(a) Below is a printout given by your computer. Fill in the blanks (____) with the appropriate numbers using the information given above. I have left a blank page after this one on which to show your work and a suggested order for doing the calculations. Give at least a brief indication, either in formulas or words, of how you got the numbers. If there is a number you can't figure out, put in a symbol for it and show how you would get all the other numbers using the symbol.

```
The regression equation is

Fans = _____ + ______Temperature


Predictor         Coef    SE Coef        T             P

Constant          ____        15      _______       1.000

Temperature       ____       ____     _______        .000


RMSE = _____   R-sq =  ____    R-sq(adj) = .6857


Analysis of Variance

SOURCE       DF    Sum Squares  Mean Squares      F        P

Regression   ___         ____        _____     _____    _____

Error        ___         ____        _____

Total        ___          560
```

(1) Find the estimated regression equation.
(2) Fill in the table below the regression equation.
(3) Fill in RMSE and MSE.
(4) Fill in the degrees of freedom in the ANOVA table, and then the rest of the ANOVA table. (Note: No calculations are needed for the p-value!)
(5) Fill in R^2.

(b) Give the units and real-world interpretations of the regression coefficients β_0 and β_1. (Note:

You do not need to quote the numbers to do this, though it may be helpful to do so if you know them.)

(c) Is there a significant linear relationship between temperature and the number of fans sold by the store? Answer this question by performing a t test. You must state the null and alternative hypotheses, both mathematically and in words, quote the p-value, and give your conclusions. You do not need to quote the test statistic and no calculations are required. Use $\alpha = .05$.

(d) Calculate a 95% confidence interval for β_0. Based on your interval, is β_0 different from 0? Explain. (Note: If you couldn't get b_0 in part (a), you may assume it is 1 for this part of the problem.) What does this interval tell you?

(e) Is temperature a good **predictor** of fan sales? Quote the number that you use to determine this and briefly explain your reasoning.

(f) What **percentage of the variability** in fan sales is explained by the regression on temperature? Quote the number that you use to

(g) A weather forecast says next week's temperature will soar to 100! Predict the number of fans you will sell next week. Suppose you want a range of possible values for the number of fans you will sell. Calculate the appropriate interval and explain your reasoning. (Use $\alpha = .05$) How many fans should you stock to be sure you have enough on hand?

Multiple Regression

(1) Harry Potter and the Sorcerer's Statistic: Polygon Pictures, the film-making branch of Mathematical Media Incorporated, is interested in knowing what factors contribute to the profitability of their movies. For their last n=27 films they have recorded Y, the box-office sales (in millions of dollars), X_1, the production costs for the film (in millions of dollars), X_2, the number of theaters in which the film was shown, X_3 the advertising budget for the film (in millions of dollars), and X_4 which is 1 if the movie featured a big name star and 0 if it didn't. They have also classified the films as action/adventure ($X_5 = 1, X_6 = 0$), comedy ($X_5 = 0, X_6 = 1$), or romance ($X_5 = 0, X_6 = 0$). A multiple regression printout for their data is shown below along with some possibly helpful statistics. Use this information to answer the following questions.

```
Correlations:
            Box Theaters      Ads
Theaters  0.900
Ads       0.925    0.889
Cost      0.950    0.912    0.927
**********************************
Summary         N     MEAN   MEDIAN    STDEV
Box            27    35.00    25.00    14.78

               MIN      MAX       Q1       Q3
```

```
Box          5.00    70.00    20.00    40.00
**********************************
The regression equation is
Box = - 0.842 + 1.84 Cost + 0.0025 Theaters - 0.628 Ads + 5.47 Star
           + 4.59 Action - 5.14 Comedy

Predictor       Coef       Stdev    t-ratio        p
Constant     -0.8423      0.9715      -0.87    0.396
Cost          1.8365      0.1339      13.71    0.000
Theaters      0.0025    0.001455       1.71    0.102
Ads          -0.6282      0.5574      -1.13    0.273
Star          5.4713      0.7191       7.61    0.000
Action        4.5880      0.6523       7.03    0.000
Comedy       -5.1441      0.6412      -8.02    0.000

RMSE = 0.8970      R-sq = 99.7%     R-sq(adj) = 99.6%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    6     5665.32      944.22   1173.54    0.000
Error        20       16.09        0.80
Total        26     5681.41
```

(a) Is the regression **overall** useful for explaining the box-office take of the movies? Justify your answer with an appropriate hypothesis test using $\alpha = .05$. You do not need to write out all the details here–just explain your basic reasoning–but on the exam you should be prepared to give the full details!

(b) What percentage of the variability in box-office take is explained by the variables in this model? What value should you use to check this and why?

(c) Does the model do a good job of predicting box-office take? Briefly justify your answer.

(d) Find a 95% confidence interval for β_4, the coefficient of the big name star variable, and give a brief real-world interpretation of your interval.

(e) Suppose Polygon Pictures gets 50% of the box office take. What is the maximum amount they could pay a big name star and be 95% (really 97.5%) sure it was economically worthwhile?

(f) Polygon Pictures is about to begin filming Harry Potter and the Sorcerer's Statistic, an adventure film that they plan to release in 3000 theaters with an advertising budget of \$2 million, production costs of \$30 million, and no big name stars. What is the predicted box office sales for this film?

(g) Suppose the Polygon CEO wants an interval which is 95% certain to contain the true box-office take of the Harry Potter film. What type of interval should she use, a confidence interval or a prediction interval? Explain briefly.

(h) What would be the difference in box-office sales if Polygon decreased the overall production costs by $1 million but added a big name star to the cast? Does this seem like a good move? Briefly justify your answer. Note that you should NOT need to make an entirely new prediction of sales.

(i) Which type of film is generally most profitable, all other things being equal? An action/adventure film, a comedy, or a romance? Explain briefly.

(2) When I Finish Summer School....I'm Going To StatisticsLand?!

Our old friend Professor Sadisticus has gone into the amusement park business. He is currently trying to determine what factors affect attendance at his parks. He has recorded Y, the number of visitors (in millions) to each of his parks each quarter for the past 5 years. Average attendance has been $\bar{Y} = 5$ or 5 million people. He has also recorded data on X_1 time (with time 1 being the first quarter, winter, five years ago), X_2 the price of tickets to the park (in dollars), X_3 the number of rides at the park, X_4 the size of the park (in acres), X_5 the population of the city in which the park is located (in hundreds of thousands of people), and X_6 the average temperature during the quarter (in degrees). He also has indicator variables for whether there were special discounts offered to local residents ($X_7 = 1$ if there was a discount and $X_7 = 0$ if there wasn't) and for the region of the country in which the park was located $X_8 = X_9 = 0$ for the west coast, $X_8 = 1, X_9 = 0$ for the midwest, and $X_8 = 0, X_9 = 1$ for the east coast.) He has fit a multiple regression of Y on these nine variables. Use the regression printout and accompanying summary statistics to answer the questions on the following pages.

```
Correlations:
               Attendance Price Rides Size
Price             -.7
Rides              .8      .5
Size               .7      .4    .9
Population         .7     -.1    .1    .2
**************************************************************************
The regression equation is

Attendance = 2 + .25Time - .2Price + .01Rides - .001Size + .05Population +
.1Temperature + 1Discount  -2Midwest - East

Predictor          Coef      SE Coef         T             P
Constant          2.000         .500      4.00         .0001
Time               .250         .100      2.50         .0142
Price             -.200         .050     -4.00         .0001
Rides              .010         .004      2.50         .0142
Size              -.001         .002     - .50         .6180
Population         .050         .025      2.00         .0480
Temperature        .100         .048      2.08         .0396
Discount          1.000         .400      2.50          0142
Midwest          -2.000        1.000     -2.00         .0480
East             -1.000        1.000     -1.00         .3196

RMSE = .200     R-Sq = 95.6%      R-Sq(adj) = 95.24%

Analysis of Variance

Source            DF          SS          MS            F          P
Regression         9       95.60       10.62        265.5      0.000
Residual Error   110        4.40         .04
Total            119      100.00
```

Note: EXCEPT WHERE INDICATED OTHERWISE YOU SHOULD USE $\alpha = .05$ FOR ALL HYPOTHESIS TESTS ON THIS PROBLEM

(a) Does the model as a whole do a good job of explaining attendance at StatisticsLand parks? Answer this question by performing an appropriate hypothesis test. State the null and alternative hypotheses both mathematically and in words, give the test statistic and p-value, and explain your real-world conclusions.

(b) Does this model do a good job of **predicting** attendance at StatisticsLand parks? Carefully justify your answer.

(c) Give the real-world interpretations of b_3, and b_7 in this model. Your answer should include the actual numerical values and units as appropriate.

(d) Based on the correlation table given above the regression printout, should X_4, the size of the park, be a good predictor of attendance? Should it's coefficient, β_4, be positive or negative? Explain briefly in each case.

(e) Are b_4, the estimated regression coefficient for X_4, and its p-value consistent with your expectations from part (e)? Justify your answer, and, if there is a lack of consistency, say what has gone wrong, giving evidence from the data to support your argument.

(f) Does it appear that there are statistically significant differences in park attendance in the three regions of the country? If so, in which region(s) is attendance the highest? Briefly justify your answers.

(g) On average, how much would you expect attendance at a theme park to decrease in a quarter if you raised the ticket prices by \$5, all other things being equal? Briefly explain your reasoning.

(h) Find the predicted attendance for the Los Seraphim theme park this summer (that is, summer of the first year after the recorded data.) Los Seraphim is a west coast city with a population of 5 million people, and an average summer temperature of 70 degrees. You may assume that the park has 50 rides, has 50 acres of space, that admission is \$45, and that there are currently no discounts being offered.

(i) Professor Sadisticus is planning to open a new theme park in the city of Hollybrick. He knows it will take a while for the park to become profitable but would like to be 95% sure that **in total** over the next 10 years attendance will be high enough so that he does not lose money on it. (i) What sort of interval should he use when predicting quarterly attendance at the park to find his projected profits over this period? (ii) Do you see any potential problems with the predictions he is making? Briefly explain your reasoning in each case. No calculations are required.

(j) Professor Sadisticus has a theory that as it gets hotter more people come to his theme parks. Because of this he is considering adding a new water ride, the Random Splatter, and setting up ice-cream stands on hot days, but before the does this he wants to be sure that his theory is correct. Perform the appropriate hypothesis test to prove Professor Sadisticus' theory. Be sure you state your null and alternative hypotheses both mathematically and in words with a justification of your choice, give the test statistic and p-value, and explain your real-world conclusions.

k Silly Sally, a summer intern at StatisticsLand, is very excited by the results of your test in part (j). She concludes that hotter weather **causes** people to visit your theme parks and proposes that new parks should be opened in desert areas like Arizona or Saharan Africa. Explain what is wrong with (i) Sally's conclusion and (ii) her proposal. Your answer should include an example of why Sally's conclusion might be wrong.