November 28th, 2011

Final Exam Practice Problems

Note: In this file are some additional practice problems for our final exam, mostly pertaining to logistic regression. I do not claim that they cover all the possible topics that are fair game for the exam. They are simply intended to supplement the various problems on the homework assignments, handouts and previous practice sets. Let me know as you procede with your studying if there is anything else on which it would be helpful to have more problems.

Logistic Regression Practice

I have started with a bunch of logistic regresison problems since we did not have a homework set on this material. You certainly don't need to do all of them! Problem 3 in particular provides a nice basic run through the key concepts. Problems 4-5 come from old exams but in classes where I had spent more time covering logistic regression. I used the printout from Problem 5 in class as an example but didn't do all of the pieces listed here. Problem 6 has a nice example of how I could work confounding issues into a logistic regression problem (part (f)).

(1) Logistic Regression Basics:

(a) Explain what the response variable is in a logistic regression and the tricks we use to convert this into a mathematical regression equation.

(b) Explain what an **odds ratio** means in logistic regression.

(c)) Explain what the coefficients in a logistic regression tell us (i) for a continuous predictor variable and (ii) for an indicator variable.

(2) Cardiovascular Disease (Based on Rosner 13.58-61): Sudden death is an important, lethal cardiovascular endpoint. Most previous studies of risk factors for sudden death have focused on men. Looking at this issue for women is important as well. For this purpose, data were used from the Framingham Heart Study. Several potential risk factors, such as age, blood pressure and cigarette smoking are of interest and need to be controlled for smilutaneously. Therefore a multiple logistic regression was fitted to these data as shown below. The response is 2-year incidence of sudden death in females without prior coronary heart disease.

Risk Factor	Regression Coefficient (b_j)	Standard Error ($se(b_j)$)
Constant	-15.3	
Blood Pressure (mm Hg)	.0019	.0070
Weight (% of study mean)	-.0060	.0100
Cholesterol (mg/100 mL)	.0056	.0029
Glucose (mg/100 mL)	.0066	.0038
Smoking (cigarettes/day)	.0069	.0199
Hematocrit (%)	.111	.049
Vital capacity (centiliters)	-.0098	.0036
Age (years)	.0686	.0225

(a) Assess the statistical significance of the individual risk factors and explain the practical implications of your findings.

(b) Give brief interpretations of the age and vital capacity coefficients.

(c) Compute the odds ratios relating the additional risk of sudden death associated with (i) a 100-centiliter decrease in vital capacity and (ii) an additional year of age after adjusting for the other risk factors.

(d) Provide 95% confidence intervals for the odds ratios in part (c)

(e) Predict the probability of sudden death for a 50 year old woman with systolic blood pressure of 120 mmHg, a relative weight of 100% a cholesterol level of 250 mg/100mL, a glucose level of 100 mg/100mL, a hematocrit of 40%, and a vital capacity of 450 centiliters who smokes 10 cigarettes per day. (Note that these numbers are near average for a healthy woman except for the cholesterol level which is high, and of course the number of cigarettes smoked.)

(3) Ear Infections (Based on Rosner 13.66): In this problem we assess the impact of two different antibiotics on the chances a child will be cured of an ear infection after adjusting for agd and whether one or both ears were infected. The variables are "Clear"–whether the infection has been cleared from both ears after 14 days treatment, "Antibiotic"–the treatment type (1 = Ceftriaxone, 0 = Amoxicillin), Age (categories under two years old, 2-5 years old and 6 year or older), and "NumEars"–the number of ears infected (either 1 or 2). STATA outputs for the pertinent logistic regression model are below. There are two versions, **logit** which gives the raw coefficients and their standard errors and **logistic** which gives the odds ratios and their standard errors.

```
. logit Clear Antibiotic NumEars TwoToFive SixPlus
Logistic regression                               Number of obs   =        203
                                                  LR chi2(4)      =      21.79
                                                  Prob > chi2     =     0.0002
Log likelihood = -129.75295                       Pseudo R2       =     0.0775
------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   .6692876   .3008256     2.22   0.026     .0796802    1.258895
     NumEars |   .0439546    .321911     0.14   0.891    -.5869793    .6748885
   TwoToFive |   1.148698   .3715113     3.09   0.002     .4205494    1.876847
     SixPlus |    1.65964   .4421503     3.75   0.000     .7930418    2.526239
       _cons |  -1.417179   .6001296    -2.36   0.018    -2.593411   -.2409466
------------------------------------------------------------------------------

. logistic Clear Antibiotic NumEars TwoToFive SixPlus
Logistic regression                               Number of obs   =        203
                                                  LR chi2(4)      =      21.79
                                                  Prob > chi2     =     0.0002
Log likelihood = -129.75295                       Pseudo R2       =     0.0775
------------------------------------------------------------------------------
       Clear | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   1.952846    .587466     2.22   0.026     1.082941    3.521528
     NumEars |   1.044935    .336376     0.14   0.891     .5560043    1.963814
   TwoToFive |   3.154084   1.171778     3.09   0.002     1.522798    6.532873
     SixPlus |    5.25742    2.32457     3.75   0.000     2.210109    12.50638
------------------------------------------------------------------------------
```

(a) Overall do these variables help explain how likely a child is to have their ear infections cleared in 14 days? Briefly justify your answer.

(b) Do these variables explain a lot the "variability" in how likely an ear infection is to clear? Explain briefly. What are the practical implications of this statement for treating ear infections in small children with antibiotics?

(c) Describe what you think would happen if you used backwards stepwise selection to find the best model for predicting whether a child's ear-infection would clear. That is, say what variables would be included in the intial model, what would happen at each step, and what you think the final model would be, and what you would have to do to verify your answer.

(d) Explain briefly how you could figure out what variable to add first in a forwards stepwise model selection procedure for this data.

(e) Which of the age categories have I used as the reference in this model?

(f) Give brief interpretations of the odds ratios for the "Antibiotic" and "TwoToFive" Variables and show how you would compute them from the information given in the first (logit) printout.

(g) Verify the calculation of the confidence interval for the coefficient of the SixPlus coefficient in the first model and show how to convert it into the confidence interval for the odds ratio given in the second printout.

(h) According to this model is their a difference in efficacy between Ceftriaxone and Amoxicillin? Write out the details of the appropriate hypothesis test using $\alpha = .05$ (hypotheses mathematically and in words, test statistic, p-value, conclusions.) Does our model show whether either antibiotic helps cure ear infections? Explain briefly.

(i) According to this model does whether one or both of a child's ears are infected affect their chance of being cured within 14 days using $\alpha = .05$? You do not need to write out the details. Just briefly jsutify your answer.

(j) After adjusting for the other factors, does age impact the likelihood of an infection clearing within 14 days? Explain briefly using $\alpha = .05$.

(k) Is there a difference in likelihood of cure between children who are 2-5 and children 6 or older? Explain briefly. (Note: You do NOT need to refit the model with a different reference group for age–the information you need is on the printouts.)

(4) Special Delivery: In the developed world most people with HIV receive some form of "highly active antiretroviral therapy" or HAART. (HAART regimens are basically cocktails of multiple drugs that are more effective because the virus is less likely to become resistant in their presence.) However in underdeveloped nations HAART is rarer because of its cost. Professor Helpful believes that HAART regimens will help reduce the risk of HIV positive pregnant women passing on the infection to their babies and must therefore be agressively promoted in poor countries. He has followed n=300 HIV positive pregant women, 100 of whom are receiving at most a basic non-HAART treatment, 100 of whom are taking HAART regimen A, and 100 of whom are taking HAART regimen B. (I'll skip the drug names to keep this simple!) He records Y, whether or not the baby is HIV positive (1 = yes, 0 = no) and which treatment regimen the mother was on ($X_1 = 1$ if the mother was on HAART A and 0 otherwise, $X_2 = 1$ if mother was on HAART B and 0 otherwise), and fits a logistic regression. The corresponding STATA printouts are below. Use them to answer the following questions.

```
. logit HIVplus HAART_A HAART_B
```

```
Logistic regression                                     Number of obs   =        300
                                                        LR chi2(2)      =       6.75
                                                        Prob > chi2     =     0.0342
Log likelihood =  -96.32681                             Pseudo R2       =     0.0339

------------------------------------------------------------------------------
     HIVplus |   Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |  -0.539        .431    -1.25   0.211    -1.383       0.305
     HAART_B |  -1.286        .534    -2.41   0.016    -2.332      -0.240
       _cons |  -1.658        .273    -6.08   0.000    -2.193      -1.124
------------------------------------------------------------------------------

. logistic HIVplus HAART_A HAART_B

Logistic regression                                     Number of obs   =        300
                                                        LR chi2(2)      =       6.75
                                                        Prob > chi2     =     0.0342
Log likelihood =  -96.32681                             Pseudo R2       =     0.0339

------------------------------------------------------------------------------
     hivplus | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |   .583      .251         -1.25   0.211       .251    1.357
     HAART_B |   .276      .147         -2.41   0.016       .097    0.787
------------------------------------------------------------------------------
```

(a) Overall, is treatment regimen useful for explaining whether a woman passes on HIV infection to her baby? Write down the mathematical hypotheses you are testing, circle the relevant p-value on one of the printouts and give your real-world conclusions using $\alpha = .05$. You do NOT need to provide any other details.

(b) Give a brief interpretation of the odds ratio for the HAART A variable and show how to compute it from the first regression printout.

(c) Do HAART A and HAART B appear to **reduce** a mother's risk of passing on HIV to her infant? Explain briefly using $\alpha = .05$ and give the p-values corresponding to the tests you are performing. You do NOT need to write out any other details of the tests.

(d) Find the odds ratio comparing the risk of HIV transmission for mothers in the HAART A group compared to those in the HAART B group. Show your work. Based on this estimate which of these treatment regimens is more effective? Briefly explain your reasoning. Do you think you can be 95% sure this treatment is better? Explain.

(5) Prenatal Care-acteristics: Professor Helpful recognizes that there are probably many factors besides treatment regimen that affect whether a mother transmits HIV to her baby. He has thus added the following variables to his logistic regression model from Question 4: X_3, the mother's viral load in copies per milliliter of blood (higher viral load is worse), X_4, the mother's age in years, X_5, the number of years the mother has been HIV positive, X_6, the number of weeks during the pregnancy for which the mother was receiving HAART therapy, and X_7 the method by which the baby was delivered (1 = C-section, 0 = natural delivery).

The new printouts are given below. Use them to answer the following questions.

```
. logit HIVplus HAART_A HAART_B VLoad Age YrsHIV WksHAART Delivery

Logistic regression                                     Number of obs   =        300
                                                        LR chi2(7)      =      32.47
                                                        Prob > chi2     =      0.000
Log likelihood =  -26.51722                             Pseudo R2       =      0.500
------------------------------------------------------------------------------
     HIVplus |   Coef.      Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |  -0.70        0.250       2.80  0.005      [-1.19,   -0.21]
     HAART_B |  -1.80        0.300       6.00  0.000      [-2.39,   -1.21]
       VLoad |0.00001    0.0000025       4.00  0.000    [.000005, .000015]
         Age |   0.10        0.050       2.00  0.046      [ 0.00,    0.20]
      YrsHIV |   0.10        0.080       1.25  0.211      [-0.06,    0.26]
    WksHAART |  -0.05        0.010      -5.00  0.000      [-0.07,   -0.03]
    Delivery |  -0.40        0.150      -2.67  0.004      [-0.69,   -0.11]
       _cons |  -5.00        0.500     -10.00  0.000      [-5.98,   -4.02]
------------------------------------------------------------------------------

. logistic HIVplus HAART_A HAART_B

Logistic regression                                     Number of obs   =        300
                                                        LR chi2(7)      =      32.47
                                                        Prob > chi2     =      0.000
Log likelihood =  -26.51722                             Pseudo R2       =      0.500
------------------------------------------------------------------------------
     HIVplus |    OddsRatio       z      P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |      0.4966       2.80    0.005        [0.3042,  0.8106]
     HAART_B |      0.1652       6.00    0.000        [0.0916,  0.2982]
       VLoad |     1.00001       4.00    0.000     [1.000005, 1.000015]
         Age |      1.1052       2.00    0.046        [1.0020,  1.2190]
      YrsHIV |      1.1052       1.25    0.211        [0.9448,  1.2928]
    WksHAART |      0.9512      -5.00    0.000        [0.9328,  0.9700]
    Delivery |      0.6703      -2.67    0.004        [0.4996,  0.8994]
------------------------------------------------------------------------------
```

(a) Find the probability that a 30 year old women on HAART A for 20 weeks of her pregnancy with a viral load of 10,000 who has been HIV positive for 10 years will have an HIV **negative** baby if she delivers by Cesarean Section. Show your work.

(b) Explain as precisely as you can the meaning of the p-value for X_7, the delivery variable. Your answer should be specific to this context and incorporate the relevant numeric value(s).

(c) (i) Give a brief interpretation of the confidence interval for the odds ratio for X_6, the weeks treated variable. (ii) Find a 95% confidence interval for the odds ratio associated with an extra MONTH (4 weeks) of HAART treatment. Based on this latter interval can you be sure that, all else equal, an extra month of HAART treatment will reduce the risk of mother to child transmission by 10%.

(d) Professor Helpful believes overfitting is an issue in this model. (i) Explain why he is correct. (ii) Give a possible real-world cause of the overfitting and say how you would check whether your idea was correct.

(iii) Say what variable you would remove first in a backwards stepwise procedure and why. (iv) What do you think would happen to the pseudo R^2 if you removed this variable? Why?

(6) Sports Fanatics: My husband, Gareth, is from New Zealand where the national sports passion is rugby (sort of like American football only better!) The national rugby team is called the All Blacks (they wear black) and their main rivals are Australia (the Wallabies) and South Africa (the Springboks). Gareth realizes that what he really cares about is whether the All Blackswin or not. Therefore he decides to perform a logistic regression with the the response variable, Y, being whether or not the All Blacks win (Y = 1 if they win and 0 if they lose). The predictors are

AB Win%=the percent of the previous ten games that the All Blacks had won going into the game in question, ranging from 0 to 100

OppWin%, (same definition for the opponents last 10 games)

Home?, an indicator variable with 1 corresponding to an All Blacks home game and 0 an away game

Temperature (the temperature at which the game was played.)

Australia? (a dummy variable with 1 corresponding to a game against archrival Australia and 0 a game against another team.)

Below are the p-value for the likelihood ratio chi-square test along with a table of coefficients, standard errors, Z scores and p-values for the various variables. Use them to answer the questions below.

```
LR chi2 p-Value < 0.0001

             Coef      SE     Z    p-value
Constant   -25.30   10.54 -2.40   0.0163
AB Win %    0.466   0.176  2.65   0.0082
Opp Win %  -0.170   0.643 -2.65   0.0081
Home?        1.45   0.660  2.20   0.0278
Temperature 0.115   0.045  2.55   0.0108
Australia? -0.245   1.890 -0.13   0.8969
```

(a) Is there evidence that at least one of the variables is a statistically significant predictor of whether the All Blacks win? Justify your answer.

(b) What does the coefficient for Temperature tell us about the relationship between Temperature and the probability that the All Blacks win? Compute the corresponding odds ratio for a 10 degree increase in temperature and explain what it means. Give a confidence interval for this odds ratio.

(c) Which variables are statistically significant? Justify your answer. Do the signs of the various coefficients make sense?

(d) Estimate the probability of the New Zealand All Blacks winning a game against South Africa played in South Africa at 50 degree temperatures where both teams have a winning percentage of 70.

(e) Find a confidence interval for the coefficient of the Home? variable and give a brief interpretation. Also find the odds ratio for the corresponding variable and a 95% confidence interval and interpret those results.

(f) The coefficient for the Home? variable seems to indicate that the All Blacks are more likely to win at home than on the road. However, somewhat surprisingly, the All Blacks turn out to win more games on the road than at home. One of my husbands MBA students (from that school on the wrong side of town) looks at these results and states that this indicates that there must be some mistake in the analysis. However, you tell them that in fact this apparent inconsistency is entirely possible even if the model is correct. Assuming that the model is correct (i.e. there are no important variables missing from the model or violations of the basic assumptions etc.) and the coefficient estimates are exactly correct how could the coefficient for Home? be positive even though the All Blacks win more games on the road?

Regression/ANOVA Problems

For regression some of the best practice problems are the warm-up problems from homework 5 and 6. You can also take the regression printouts from the midterm 1 practice set and use them to think about the topics we have covered more recently. I have included below an extension about one of those problems which I had previously included only in ANOVA form and added some regression parts. This is the problem as it originally appeared in an exam. I've also taken a problem from the Midterm 1 practice that had a good example of an outlier and added some additional parts to it. Let me know if after going through all those options there is anything on which you feel short of practice....

(1) Analysis of Varying Medications: A researcer is analyzing methods of reducing cholesterol levels. She is interested in the relative merits of diets versus cholesterol lowering medications. For each of 65 subjects who began the study with high cholesterol she records total blood cholesterol level (in mg per deciliter) after 6 months participation in the study. The patients are divided into G=5 groups: a control group (C) which receives a placebo, a vegetarian diet group (V), a low fat diet group (LF), a low dose medication group (LD) and a high dose medication group (HD). STATA printouts below show the group means, standard deviations, and group sizes, along with an ANOVA table which seems to be missing a few numbers. Use this information to answer the questions on the following pages.

```
. oneway Cholesterol Group, tabulate

           |       Summary of Cholesterol
           |
     Group |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         C |       240        1.22           25
         V |       225        1.18           10
        LF |       230        1.10           10
        LD |       215        1.02           10
        HD |       200        1.11           10
-----------+------------------------------------
     Total |       226.2     13.51           65

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups        ______      ___      _____     _____      0.000
 Within groups        ______       60      80.00
------------------------------------------------------------------------
    Total            11681.4      ___
```

(a) Fill in the missing entries (marked ____) in the above tables. You do not need to give your reasoning though it may help if you make mistakes. There is an easy way and a hard way to do this! Try to do it the easy way! Note that you will be able to do the rest of the problem even if you can not do part (a).

(b) Based on this data is there evidence that any of the group means are different from each other? Justify your answer by performing an appropriate hypothesis test. Be sure to state the null and alternative hypotheses, both mathematically and in words, give the p-value, and your real-world conclusions.

(c) Suppose that instead of doing an ANOVA we had fit a regresison model to this data using the vegetarian diet group as the reference group. Write down the estimated regression equation we would have obtained.

(d) What percentage of the variability in cholesterol levels is explained by the treatment group to which a subject belonged? Show your calculations or explain your reasoning.

Below is a table showing test statistics and p-values for pairwise comparisons of the different group means for this ANOVA. Use it to help answer parts (e)-(f).

Pair	t_{obs}	p-value	Pair	t_{obs}	p-value
C vs V	4.48	.00003	C vs LF	2.99	.00404
C vs LD	7.47	.00000	C vs HD	11.95	.00000
V vs LF			V vs LD	2.50	.01517
V vs HD	6.25	.00000	LF vs LD	3.75	.00040
LF vs HD	7.50	.00000	LD vs HD	3.75	.00040

(e) The test comparing the vegetarian diet group to the low fat diet group is missing. State the null and alternative hypotheses mathematically and in words, compute the test statistic and an approximate p-value and explain your real-world conclusions. (Note: Make sure you carefully show your calculation of the standard error.)

(f) Which pairs of means are significantly different from one another at the $\alpha = .05$ level without adjusting for multiple testing? Explain briefly.

(g) According to the Bonferroni method, what significance level should you use for the individual tests for differences of means to get an overall significance level of $\alpha = .05$? Explain briefly. Use your answer to repeat part (f), adjusting for multiple comparisons. Indicate any results that have changed.

(h) The researcher is interested in comparing the average cholesterol level of people in the two diet groups with that of the people in the low dose medication group. Write down an appropriate linear combination, L, for the comparison she wishes to do. Give your best estimate of L and the corresponding standard error and use these numbers to find a 95% confidence interval for L. Give a brief interpretation of your interval and explain whether the researcher can conclude there is a difference in efficacy between diets and the low dose medication in reducing cholesterol levels.

Obviously there are factors other than treatment group which could affect a person's cholesterol level. Thus, the researcher has fit a multiple regression of cholesterol level on treatment group, age, weight and whether or not the person has a family history of coronary artery disease (1 = yes and 0=no). Use the STATA multiple regression printout to answer the remaining parts of the question.

```
. reg Birthweight EX LOW HIGH

      Source |       SS       df       MS              Number of obs =      65
-------------+------------------------------           F( 7 , 57  )  =    73.3
       Model | 10513.3         7   1501.9              Prob > F      =   0.000
    Residual |  1168.1        57     20.5              R-squared     =   0.900
-------------+------------------------------           Adj R-squared =   0.888
       Total | 11681.4        64                       Root MSE      =   4.528

------------------------------------------------------------------------------
 Birthweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |       0.5        0.2        2.50   0.0153      0.10       0.90
      WEIGHT |       0.6        0.2        3.00   0.0040      0.20       1.00
     HISTORY |      30.0        5.0        5.00   0.0000     20.00      40.00
         VEG |      -2.0        1.6       -1.25   0.2164     -5.20       1.20
     LOW FAT |       1.0        0.8        1.20   0.2351     -0.60       2.60
    LOW DOSE |      -5.0        3.0       -1.67   0.1004    -11.00       1.00
   HIGH DOSE |     -25.0        5.0       -5.00   0.0000    -35.00     -15.00
       _cons |     100.0       25.0        4.00   0.0002     50.00     150.00
------------------------------------------------------------------------------
```

(i) In terms of percentage of variability explained and accuracy of predictions does this model do a better job than the simple ANOVA from parts (a)-(h)? Explain briefly what numbers from the printout you are looking at to answer this question and also perform an appropriate hypothesis test. Does this model make good predictions? Explain.

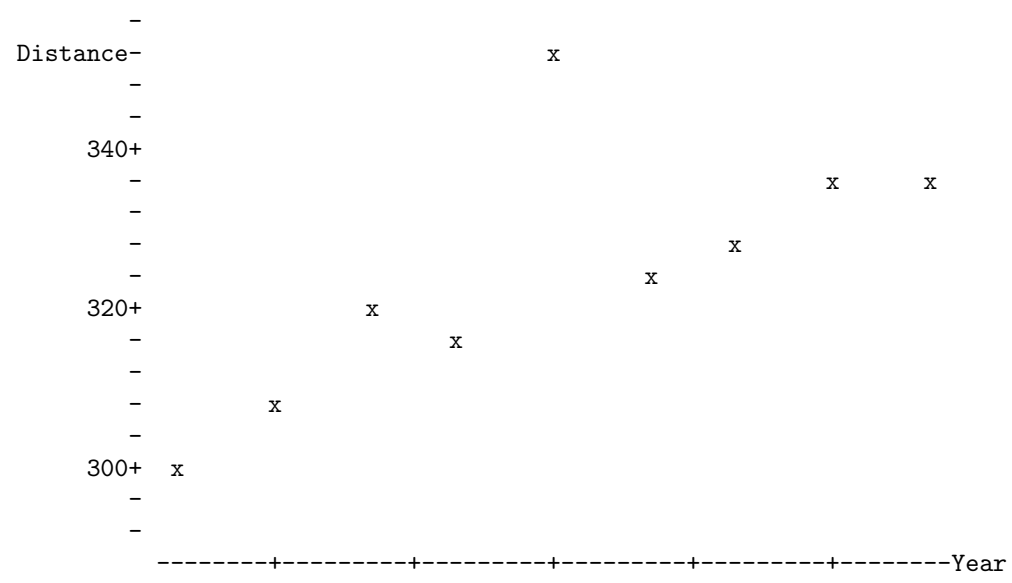
(j) After adjusting for age, weight, and family history, does it appear that the diets or medication doses have a significant impact on cholesterol levels compared to the control group? Briefly justify your answer.

(k) Your answer to part (j) is different from what you found in parts (f) and (g). Explain what has happened and what it implies about whether the researcher performed a properly randomized study.

(2) 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
------------------------------------------------------------------------------

. reg Distance Year Yearsq

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  2,     6) =    6.88
       Model |   1394.3493     2  697.174648           Prob > F      =  0.0280
    Residual |  608.150703     6  101.358451           R-squared     =  0.6963
-------------+------------------------------           Adj R-squared =  0.5951
       Total |      2002.5     8    250.3125           Root MSE      =  10.068

------------------------------------------------------------------------------
```

Distance	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Year	225.7233	141.1208	1.60	0.161	-119.5868	571.0333
Yearsq	-.0570718	.0358538	-1.59	0.163	-.1448028	.0306591
_cons	-222852.3	138860.1	-1.60	0.160	-562630.8	116926.1

****

(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. Explain what diagnostics you could use to determine whether this point is an outlier or an influential point and what each one would tell you. Intuitively do you expect the point to be highly influential? Does it appear to have high leverage? Is it an outlier? Explain. What would happen to your answers to (b)-(d) if this point were removed?

(f) Use STATA to get the residual, histogram, and normal quantile plots for the simple linear regression (or if you don't want to take the time to do so just look at the scatterplot above–on the exam obviously I would just give them to you.) Does it appear that any of our regression assumptions have been violated? Make sure you state each of the assumptions that can be checked with each plot and whether you think they are OK. What do you think is causing any problems you see, and how might you fix them?

(g) A zealous sports fan suggests that the winning distance in the long jump cannot increase for ever, but should instead level off. He therefore suggests fitting a curvilinear regression to the data. The second printout shows the results of fitting the model

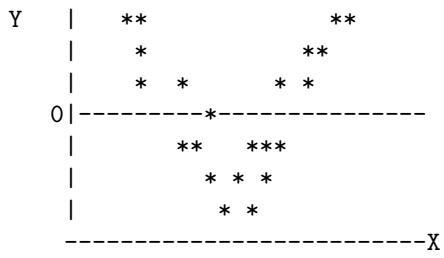
$$Y = \beta_0 + \beta_1 Year + \beta_2 Year^2$$

Is it worth adding the term $Year^2$ to the model based on the data presented here? Answer this question using an appropriate test. Make sure you state the null and alternative hypotheses, the p-value for the test, and your conclusions. Is this likely to introduce multicollinearity to the model? Explain why it might, how you could check, and what you could do to fix the problem if it exists. Try it and see if it helps. Finally, in real-world terms is the quadratic model likely to be completely appropriate for this data? Can you suggest an alternate transformation that might be better? Explain.

(3) Regression Assumptions

A residual plot from a simple linear regression analysis is shown below. It is followed by four statements about the error assumptions for this model. In each case, say whether the statement is correct. If the

statement is not correct, give an appropriate statement about the error assumption referred to.



(a) The mean 0 assumption is correct because there are approximately as many residuals above the line as below it.

(b) The constant variance assumption is violated because there is a curved pattern to the data.

(c) The errors for this data set are approximately normally distributed.

(d) A linear model is not appropriate for this data set because of the curved pattern in the data.