Final Exam

Tuesday the 7th of December, 2010

Name:_____

General Comments:

- This exam is closed book. However, you may use four 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 page. There are no tables attached at the end because you will not need them for this test. Make sure you have all 18 pages!

- The exam is 180 minutes long. There are 3 questions, worth a total of 100 points. They are not equally weighted, nor are they of equal difficulty. The number of points each question is worth is printed with the problem. Read the questions carefully. If you are unsure of the interpretation, come ask.

- **You must show your work** 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 a good idea to explain your reasoning **briefly** 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. GOOD LUCK!

Question Number	Total Points Possible	Score Received
1	40	
2	25	
3	35	
Total	100	

THE STORY BEHIND THE EXAM: Dr. Urtha Green, the environmental health scientist from the midterm is back! She has been asked to help the city of Los Seraphim evaluate its pollution levels and how they relate to asthma rates and symptoms among the city's children. During the exam you will help her analyze some of the relevant data and interpret the results.

Question 1: Smog and the City (40 points, 65 minutes)

First Dr. Green will be analyzing the annual pollution levels in Los Seraphim over the past several decades. She has been given average pollution levels, Y, in thousands of particles per cubic centimeter from 1975-2010. She decides to let X be a variable representing time in years since the start of the study period (so $X = 0$ for 1975 and $X = 35$ for 2010). She starts her analysis by fitting a simple linear regression of Y on X.

```
. regress pollution year

      Source |       SS       df       MS              Number of obs =      36
-------------+------------------------------           F(  1,    34) =   65.82
       Model |  1735.45291     1  1735.45291           Prob > F      =  0.0000
    Residual |  896.410877    34  26.3650258           R-squared     =  0.6594
-------------+------------------------------           Adj R-squared =  0.6494
       Total |  2631.86378    35  75.1961081           Root MSE      =  5.1347

------------------------------------------------------------------------------
   pollution |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   .6683607   .0823794     8.11   0.000     .5009456    .8357758
       _cons |   18.08226    1.67651    10.79   0.000     14.67518    21.48933
------------------------------------------------------------------------------
```

Part a

Is there a significant **positive** linear relationship between pollution level and year? Briefly justify your answer using $\alpha = .05$. You do not need to give all the details of the test but do give the mathematical version of the alternative hypothesis.

Part b

Next Dr. Green obtains a scatterplot, residual plot, histogram and qq-plot for the model from part (a). Use the plots to discuss whether each of the main regression assumptions is violated. You should make it clear which plot(s) you are using to assess each assumption. How do your conclusions fit with your answer to part (a)?

Inspired by your answer to part (b) Dr. Green decides to fit a quadratic model to the annual pollution data. The STATA printout for her model is shown below.

```
. regress pollution year yearsq

      Source |       SS       df       MS              Number of obs =      36
-------------+------------------------------           F(  2,    33) =   97.16
       Model |  2249.78989     2  1124.89495            Prob > F      =  0.0000
    Residual |  382.073892    33  11.5779967            R-squared     =  0.8548
-------------+------------------------------           Adj R-squared =  0.8460
       Total |  2631.86378    35  75.1961081            Root MSE      =  3.4026

------------------------------------------------------------------------------
   pollution |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |    2.04054   .2129899     9.58   0.000     1.607209    2.473871
      yearsq |  -.0392051   .0058821    -6.67   0.000    -.0511724   -.0272378
       _cons |   10.30657   1.610995     6.40   0.000     7.028981    13.58417
------------------------------------------------------------------------------

    Variable |       VIF       1/VIF
-------------+----------------------
        year |     15.22    0.065694
      yearsq |     15.22    0.065694
-------------+----------------------
    Mean VIF |     15.22

. corr year yearsq
(obs = 36)
             |     year   yearsq
-------------+------------------
        year |   1.0000
      yearsq |   0.9666   1.0000
```

Part c

Is the quadratic model a significant improvement over the simple linear model? Justify your answer by performing an appropriate hypothesis test using $\alpha = .05$. State the null and alternative hypothesis mathematically and in words and give your conclusions. Also explain briefly what the negative sign on the year-squared variable tells you in real-world terms.

Part d

Is there evidence of significant multicollinearity in the quadratic model? Briefly explain your reasoning and describe how you could correct the problem if there is one.

Now Dr. Green learns that in the year 2000 the city of Los Seraphim passed new environmental rules aimed at reducing pollution. City administrators would like to know if the laws have had an impact. Therefore Dr. Green decides to try yet another version of her model. She uses as predictors her original year variable X, an indicator variable, "envlaw" for whether the environmental laws were in effect which is 1 if the year is 2000 or later (i.e. if $X \geq 25$) and 0 otherwise, and an interaction, "year_envlaw" between X and the indicator variable. The printout for her new model is shown below. Use it to answer the remaining parts of the problem.

```
. regress pollution year envlaw year_envlaw

      Source |       SS       df       MS              Number of obs =      36
-------------+------------------------------           F(  3,    32) =  136.38
       Model |  2440.94775     3  813.649248           Prob > F      =  0.0000
    Residual |  190.916038    32  5.96612618           R-squared     =  0.9275
-------------+------------------------------           Adj R-squared =  0.9207
       Total |  2631.86378    35  75.1961081           Root MSE      =  2.4426

------------------------------------------------------------------------------
   pollution |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        year |   1.119491   .0677446    16.53   0.000     .9814999    1.257482
      envlaw |   58.55486   7.089121     8.26   0.000     44.11479    72.99493
 year_envlaw |  -2.314979   .2425424    -9.54   0.000    -2.809021   -1.820936
       _cons |   13.51635   .9484241    14.25   0.000     11.58447    15.44823
------------------------------------------------------------------------------
```

Part e

Write down the equations describing the relationship between the pollution level and year (X) for (i) years before 2000 when the environmental laws were not yet in effect and (ii) for the years 2000 and later when the new rules had taken effect. Use these equations to draw a rough sketch of the estimated relationship between Y and X over the period 1975-2010 and say whether the form of the relationship makes real-world sense. (Hint: This involves writing down the full estimated regression model from the printout and plugging in appropriate values for the indicator variable. For the sketch you may find it helpful to get the predicted values in 1975, 2000 and 2010).)

Part f

Is the interaction between year, X, and the indicator for whether the new environmental laws were in effect statistically significant? Briefly justify your answer. Based on this and your answer to part (e) do the laws seem to have had a meaningful impact? Explain carefully.

Part g

From the plots earlier in the problem, Dr. Green is concerned that there may be some unusual points in the data set. She has therefore calculated some outlier diagnostics. These are shown below for the points with the most extreme values, ordered in terms of the absolute value of the studentized residual. Which year the point belongs to is shown in the first column. For comparison purposes the next highest value of each diagnostic is given in the last row. Discuss briefly whether any of these points are a concern using the thresholds we learned in class and say what a real-world explanation for the worst point might be.

Year	StudentizedRes	leverage	CooksDistance
1975	-4.07	.15	.49
1980	2.38	.08	.10
1978	2.25	.10	.13
2000	0.76	.32	.07
2010	0.22	.32	.006
Next Highest	-1.41	.23	.07

Part h

Explain briefly what you would expect to happen to $R^2, RMSE$ and b_1 if the worst outlier were removed from the model.

Part i (Optional Bonus)

Why do you think the leverage for the Year 2000 point is so high when it is in the middle of the data set? Is this a problem? Discuss briefly.

Question 2: Asthma As Math (25 points, 40 minutes)

In addition to studying pollution patterns, Dr. Green is interested in the role of air quality as a risk factor for childhood asthma. She has participated in a study that followed 1000 children from birth to age 10, recording whether or not they developed asthma during that period ($Y = 1$ for yes and $Y = 0$ for no). The study also collected information on a potential risk factors and protective effects from the child's first year of life including X_1, an indicator for whether the child's family lived in an urban setting (Yes = 1, No = 0), X_2, the average annual pollution level in thousands of particles per cm^3 for the county in which the child lived, X_3, an index of socio-economic status for the child's family (higher is better),X_4, the number of months for which the child was breast-fed, X_5, an indicator for whether there was a family history of asthma (1 = Yes, 0 = No), and X_6, gender (1 = Female, 0 = Male). Printouts for some preliminary models are shown below.

```
logit asthma urban                                Number of obs   =       1000
                                                  LR chi2(1)      =      10.60
                                                  Prob > chi2     =     0.0011
Log likelihood = -367.55872                       Pseudo R2       =     0.0142
------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   .6742532   .2147084     3.14   0.002     .2534326    1.095074
       _cons |  -2.408854   .1817355   -13.25   0.000    -2.765049   -2.052659
------------------------------------------------------------------------------
*******************************************************************************
logit asthma pollution                            Number of obs   =       1000
                                                  LR chi2(1)      =      24.54
                                                  Prob > chi2     =     0.0000
Log likelihood =  -360.5918                       Pseudo R2       =     0.0329
------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   pollution |   .0682789   .0141742     4.82   0.000     .0404979    .0960598
       _cons |  -3.487279   .3476386   -10.03   0.000    -4.168638   -2.805919
------------------------------------------------------------------------------
*******************************************************************************
logit asthma urban pollution                      Number of obs   =       1000
                                                  LR chi2(2)      =      24.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -360.57094                       Pseudo R2       =     0.0330
------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |  -.0597649   .2923861    -0.20   0.838    -.6328312    .5133013
   pollution |   .0709072   .0191487     3.70   0.000     .0333765    .1084379
       _cons |  -3.506273   .3596161    -9.75   0.000    -4.211108   -2.801439
------------------------------------------------------------------------------
```

Part a

Dr. Green believes that living in an urban setting and being exposed to higher levels of pollution are both individually associated with increased risk of developing childhood asthma. Do her initial models support these theories? Explain briefly using $\alpha = .05$.

Part b

Dr. Green believes that the reason urban setting is a risk factor is primarily that urban settings have higher pollution levels. To see whether the data match her theory she fits a model with both the urban setting indicator and pollution level as predictors of whether or not a child gets asthma. Explain (i) Why Dr. Green's theory is one of mediation and (ii) whether her data, so far as you have them, are in fact consistent with the idea of either full or partial mediation and (iii) what additional tests you would want to do to complete part (ii), being as explicit as you can about what technique(s) we have learned could be used to carry out those tests.

Now Dr. Green adds in the other variables from the study and fits a new logistic regression model. The printout is shown below. Use it to answer the remaining parts of this problem.

```
. logit asthma urban pollution ses breastfed famhist gender

Logistic regression                                     Number of obs   =       1000
                                                        LR chi2(6)      =      82.04
                                                        Prob > chi2     =     0.0000
Log likelihood = -331.83839                             Pseudo R2       =     0.1100

------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |  -.0771939   .3020666    -0.26   0.798    -.6692335    .5148458
   pollution |   .0766367   .0200131     3.83   0.000     .0374118    .1158616
         ses |  -.0818089   .0345813    -2.37   0.018     -.149587   -.0140308
   breastfed |    .005728   .0103221     0.55   0.579    -.0145029    .0259589
     famhist |   1.180125   .2230401     5.29   0.000     .7429742    1.617275
      gender |  _________   .2178808    -5.02   0.000    -1.521761   -.6676835
       _cons |  -.0470419   1.655807    -0.03   0.977    -3.292364     3.19828
------------------------------------------------------------------------------

. logistic asthma urban pollution ses breastfed famhist gender

Logistic regression                                     Number of obs   =       1000
                                                        LR chi2(6)      =      82.04
                                                        Prob > chi2     =     0.0000
Log likelihood = -331.83839                             Pseudo R2       =     0.1100

------------------------------------------------------------------------------
      asthma | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   .9257104   .2796262    -0.26   0.798     .5121009     1.67338
   pollution |    1.07965   .0216071     3.83   0.000      1.03812     1.12284
         ses |    .921448   .0318649    -2.37   0.018     .8610635    .9860672
   breastfed |   1.005744   .0103814     0.55   0.579     .9856018    1.026299
     famhist |    3.25478   .7259464     5.29   0.000     2.102178    5.039341
      gender |   .3346326     .07291    -5.02   0.000     .2183271    .5128953
------------------------------------------------------------------------------
```

Part c

Which conditions appear to be risk factors for and which appear to be protective against developing childhood asthma? Briefly justify your answer.

Part d

Give brief interpretations of the odds ratio estimates for the family history and SES variables. Also show how to calculate the estimated regression coefficient for the gender variable which seems to be missing from the first printout.

Part e

In the city of Los Seraphim in 2008 the average pollution level was 35 thousand particles per cm^3. Find the predicted probability of developing asthma for a boy born during that year to parents who had asthma and an SES index score of 50 and who was breastfed for 6 months.

Part f (Optional Bonus)

The boy in part (e) has a little sister due next year (2011). Assuming the family hasn't moved, the air pollution levels in Los Seraphim follows the pattern predicted by the model in Problem 1 and that the girl is also breastfed for 6 months, find the odds ratio for her risk of developing asthma compared to her brother. You should do this WITHOUT calculating the full log odds for the girl.

Question 3: Wheezy Does It (35 points, 50 minutes)

In addition to being a risk factor for asthma, pollution can increase asthma symptoms. Dr. Green has conducted a study in Los Seraphim of 100 children with asthma. Each child is assessed on a randomly selected day and an asthma symptom score computed. A score of 0 corresponds to no symptoms while a score of 100 correpsonds to symptoms requiring hospitalization. Dr. Green has also recorded various factors on the day of the interview including season (winter, spring, summer, fall), temperature (in degrees Fahrenheit), humidity (as a percentage), barometric pressure (in inches of mercury), pollution level (in thousands of particles per cm^3), and whether or not the child has allergies. She starts by fitting an ANOVA to see whether asthma symptoms vary by season.

```
. oneway symptoms season, tabulate

           |         Summary of symptoms
    season |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
    Winter |        30.0        20.5          25
    Spring |        37.5        24.5          25
    Summer |        28.3        24.3          25
      Fall |        24.1        21.3          25
-----------+------------------------------------
     Total |        30.0        22.9         100

                       Analysis of Variance
   Source              SS         df       MS         F     Prob > F
------------------------------------------------------------------------
Between groups       2367.1         3     789.0     1.53     0.2113
 Within groups      49456.7        96     515.2
------------------------------------------------------------------------
    Total           51823.8        99     523.5
************************************************************************
Pairwise comparisons: Note-these are UNADJUSTED p-values

          Spring   Summer  Fall
Winter     .242    .793   .360
Spring             .153   .039
Summer                    .513
```

Part a

Is there overall evidence that asthma symptoms vary by season at $\alpha = .05$? Explain briefly, including as part of your answer the mathematical hypotheses for a classical ANOVA.

Part b

Suppose Dr. Green had fit the ANOVA as a regression model with Fall as the reference season. What would the estimated regression equation have been? Show your work.

Part c

Which, if any, of the pairwise comparisons of seasonal differences in the second part of the printout survive a Bonferroni correction for multiple comparisons? Does this fit with your answer to part (a)? Explain briefly.

Now Dr. Green fits a multiple regression adding all the primary predictors. She uses dummy variables for season, with Fall as the reference, and an indicator for whether the child has allergies defined so that 1 = Yes and 0 = No. Her STATA printout is shown below.

```
. regress symptoms winter spring summer temperature humidity pressure pollution
allergies

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  8,    91) =  498.53
       Model |  50667.6642     8  6333.45802           Prob > F      =  0.0000
    Residual |   1156.0867    91  12.7042495           R-squared     =  0.9777
-------------+------------------------------           Adj R-squared =  0.9757
       Total |  51823.7509    99  523.472231           Root MSE      =  3.5643

------------------------------------------------------------------------------
    symptoms |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      winter |  -1.939383   1.685963    -1.15   0.253    -5.288342    1.409576
      spring |    2.71738   1.114672     2.44   0.017     .5032212    4.931539
      summer |  -3.064418   1.427374    -2.15   0.034    -5.899722   -.2291146
 temperature |  -.3377354   .0798993    -4.23   0.000    -.4964456   -.1790252
    humidity |   .2631626   .0683126     3.85   0.000     .1274679    .3988573
    pressure |   .0692811   .1194079     0.58   0.563    -.1679081    .3064703
   pollution |   2.282227   .0853439    26.74   0.000     2.112702    2.451752
   allergies |   38.70296   .7161652    54.04   0.000     37.28038    40.12553
       _cons |  -44.06224   7.623552    -5.78   0.000    -59.20549   -28.91899
------------------------------------------------------------------------------
. test winter = summer
       F(  1,    91) =    0.30
            Prob > F =    0.5872

. test winter = spring
       F(  1,    91) =    7.46
            Prob > F =    0.0076

. test summer = spring
       F(  1,    91) =   15.31
            Prob > F =    0.0002
```

Part d

Have any of the seasonal differences changed as to significance, sign, or magnitude? What does this tell you about the variables you have added to the model and does this make real-world sense?

Part e

Does this model explain a high percentage of the variability in asthma symptoms? Does it do a good job of predicting asthma symptoms? Briefly justify your answer in each case.

Now Dr. Green decides to add interaction terms between whether the child has allergies and each of pollution and humidity. Her new printout is shown below.

```
regress symptoms2 winter spring summer temperature humidity pressure pollution
allergies allhum allpoll
```

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F( 10,    89) =  771.65
       Model |  51232.8468    10  5123.28468           Prob > F      =  0.0000
    Residual |  590.904096    89  6.63937186           R-squared     =  0.9886
-------------+------------------------------           Adj R-squared =  0.9873
       Total |  51823.7509    99  523.472231           Root MSE      =  2.5767
------------------------------------------------------------------------------
   symptoms2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      winter |  -1.064733   1.226077    -0.87   0.388    -3.500922    1.371456
      spring |    3.14173   .8114112     3.87   0.000     1.529473    4.753987
      summer |  -2.874877   1.040811    -2.76   0.007    -4.942947   -.8068069
 temperature |  -.3183712   .0578776    -5.50   0.000    -.4333727   -.2033697
    humidity |   .0483028   .0600492     0.80   0.423    -.0710137    .1676192
    pressure |   .0883956   .0867499     1.02   0.311    -.0839747    .2607659
   pollution |   2.085843   .0863075    24.17   0.000     1.914352    2.257335
   allergies |   10.39306   3.766892     2.76   0.007     2.908323    17.87779
      allhum |   .4845818      .0628     7.72   0.000     .3597996     .609364
     allpoll |   .4880411   .1156469     4.22   0.000     .2582532    .7178291
       _cons |  -34.37588   5.811376    -5.92   0.000    -45.92296    -22.8288
------------------------------------------------------------------------------
```

Part f

Is the model with the interactions a significant improvement over the previous model? Justify your answer by performing an appropriate test. State your null and alternative hypotheses mathematically and in words and explain your real-world conclusions. (You will not be able to get the exact p-value but your experience with other tests should tell you whether the result is significant.)

Part g

Explain briefly in real-world terms what the model tells you about the joint effects of allergies, humidity and pollution levels on asthma symptoms. In particular does it seem that humidity and pollution are important factors for all children?

Part h

If you were performing a manual backwards stepwise model selection procedure, what variable, if any would you choose to remove first? Explain your reasoning carefully.

Part i (Optional Bonus)

The answer to part (f) may have seemed obvious given the p-values of the two interaction terms. Can you think of a situation in which the new variable(s) you added to a model were all significant but the new model was NOT a significant improvement over the old model? Discuss briefly.

CONGRATULATIONS!! YOU ARE DONE. HAVE A GREAT WINTER HOLIDAY!