201A Winter 2010 Final Exam With Solutions

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.

Solution: Yes, there is a significant positive linear relationship between year and pollution. The p-value for the two-sided test for the year variable is 0. We actually want a 1-sided test here so we would divide the p-value by 2 but since even the 2-sided p-value is highly significant we don't really

need to worry about that. To determine that the relationship is positive we note that the estimate of the year coefficient is $b_1 = .668$ which is certainly positive–and indeed the whole confidence interval for β_1 lies above 0. Whichever way we look at this there is a significant positive relationship.

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)?

Solution: The four assumptions we make about the errors are that they are mean 0, independent, constant variance and normally distributed. The first three assumptions are checked with a residual plot. We need the errors to be centered about 0 for all values of X, not show any systematic shape pattern that suggests we had the wrong shaped model and for the band of errors to be the same width at each X. here we have a clear curved pattern with errors mostly below 0 in years 0-10 and 30+ and mostly above 0 in years 10-30. Thus the mean 0 and independence assumptions are violated. It looks like maybe we should have used a quadratic (parabola) model or something similar. The constant variance assumption is a little harder to judge. There is some suggestion that the errors are more spread out in the first 10 years and that after that the band gets narrower. We accepted any reasonable answer on this that made it clear you knew what you were looking for. The normality assumption can be checked with a histogram or qq plot. Here the histogram is mostly hump-shaped although there is a little bit of a suggestion of skew on the low end. The qq plot is moderately straight but there is a bit of curvature away fro the line, especially at the low end. Overall the normality assumptions looks fairly good but certainly not perfect.

In part (a) we said there was a significant linear relationship between years and pollution. Here the residual plot suggests that a linear model is not appropriate for these data. At first this may seem like a contradiction. However all we really said in part (a) was that years helped predict pollution–it was better than NOT using years. We did not show that the linear model was the **best** model. The plots in part (b) suggest that a curved model would be **even better** than the already significant linear model. Just because a model is significant doesn't mean it has to be the best one!!

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.

Solution: Yo find out if the quadratic model is a significant improvement over the linear model we need to check whether or not the years squared term is significant. If $\beta_2 = 0$ then years squared does not add anything to the model which would mean that the linear model was just as good. Our hypotheses are:

$H_0 : \beta_2 = 0$–the quadratic model is no better than the linear model.
$H_A : \beta_2 \neq 0$–the year squared term is worth adding to the model, meaning the curvilinear shape describes the data better than the linear shape.

Our test statistic is $t_{obs} = -6.67$ and the p-value is 0 which means that the year squared term is highly significant. The quadratic model fits better than the linear model.

The negative sign on the year squared term means that this is a downward-opening parabola. At first the pollution levels in this city were going up. However at a certain point they peaked and started going down again. This would make sense if the city started doing something to improve it's air quality.

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.

Solution: There is lots of evidence of multicollinearity between years and years squared. First, the variance inflation factors for these variables are huge at over 15 (our rough rule of thumb was that 4 was high and 10 was very high). Correpsondingly the correlation between years and years squared is .97 which is extremely high. This is an example of what we called a "structural" multicollinearity–it is the result of the way we defined our X and X^2 variables. We could fix the problem by centering, namely using $X - \bar{X}$ and $(X - \bar{X})^2$ as our predictors. This would result in the same predicted values but the correlation between the two variables would be much smaller.

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).)

Solution: First we get the equation for the years before 2000. In this case both the indicator for the environmental law and the interaction term are 0. The equation becomes

$$\hat{Y} = 13.52 + 1.12Year$$

For the years after 2000 the indicator for the environmental law is equal to 1 and we get

$$\hat{Y} = 13.52 + 1.12Year + 58.55 - 2.315 * 1 * Year = 72.07 - 1.195Year$$

For the sketch we note that data collection began in 1975 (Year = 0) and that the first equation applies up until 2000 (i.e. for Year < 25) while the second line applies for 2000-2010 (i.e. for Year ≥ 25). Each piece of the equation is a straight line, they just have different slopes. Thus if we get the values at $Year = 0, 25$ and 35 we can just connect them with straight lines. When $Year = 0$ our predicted value is just $\hat{Y} = 13.52$. When $X = 25$ we switch to the second equation. We get $\hat{Y} = 72.07 - 1.195(25) = 42.195$. Finally, when $X = 35$ we get $\hat{Y} = 72.05 - 1.195(35) = 30.245$. Our model thus predicts that pollution levels increased steadily from about 13.5 in 1975 to a high of to 42.2 and then started going back down again. This makes perfect sense if in fact the environmental regulations are effective–they should gradually start to make the pollution levels go down.

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.

Solution: To determine whether the interaction is significant we simply look at the p-value for the year*envlaw term. Since this p-value is 0 the interaction term IS significant. This means that the way in which pollution levels were changing over time DID depend on whether or not the environmental law was in place. This strongly suggests that the environmental law did have an effect–in particular the pollution levels began to decrease after the law was passed since the coefficient of the interaction was negative. However we can't be absolutely sure the effect we have seen is causal. There could have been something else that happened around that time that impacted the pollution levels (for instance the city could have started getting smaller, meaning there would be fewer cars on the road, or manufacturing companies could have left meaning there would less pollution from them.)

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

Solution: By far the worst point is the one in the year 1975. It's studentized residual of 4 is well above the usual thresholds of 2 or 3. The 1980 and 1978 points are borderline by this criteria but are not nearly as bad as the one from 1975. The 1975 point also has a much higher Cook's distance at .49 than any of the other points, meaning it is by far the most influential. Recall that Cook's Distance measures the overall impact of the point on the fitted values. Our textbook suggested that values above 1 are fairly high. While the 1975 point doesn't rise to that standard it is much much higher than any of the other points and so is worth a bit of attention. Our usual rule of thumb for leverage was that values above $2(p+1)/n$ were high where p is the number of predictors and n is the sample size. Here $p = 3$ and $n = 36$ so our cutoff would be $8/36 = .22$. The points in years 2000 and 2010 are a bit high by this standard but since their Cook's Distances are very small this is probably not a concern. Overall the 1975 point is the only one of real concern and even it is not too bad.

In 1975 it looks like the pollution levels were much lower than would have been expected according to our model. There are a couple of possible explanations. Possibly there was a lot of rain that year (rain tends to knock pollutants out of the air.) Perhaps some large manufacturers moved into the city after 1975 which made the pollution levels rise. We accepted any reasonable explanation as long as it was "real-world" and not just something statistical like random variation.

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.

Solution: The 1975 point doesn't fit our model very well. Therefore if we remove it we'd expect the model fit to improve and this R^2 should go up and RMSE should go down. For the initial slope, b_1, we note that the 1975 point probably pulls the end of the line down towards itself. If we remove it, the line will be able to start out at a higher level. However the later points won't really change. As a result the slope will be **smaller** as the fitted line will be **flatter.**

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.

Solution: High leverage occurs when a point is at the "edge" of the data set with respect to the X values. The 2000 point is the change point for the interaction term. Essentially we are fitting two separate linear models–one before 2000 and one after 2000. Thus the 2000 point is at the edge for both pieces. This is not particularly a problem as nothing unusual has happened in 2000 and indeed our diagnostics show that the 2000 point is not particularly influential.

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 $\alpha = .05$.
Solution: The data do support her theory. Both urban (p-value = .002 in the first printout) and pollution level (p-value = .000 in the second printout) are highly significant individual predictors of whether or not a child has asthma.

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.
Solution: A variable, Z, is a mediator of the relationship between X and Y if X is related to Y because X causes Z which in turn causes Y. In other words, X affects Y indirectly through the mediator, Z. Here Dr. Green believes that urban settings are a risk factor for asthma because pollution causes asthma and cities tend to have high pollution levels. This is exactly the mediation set up. In order to have mediation we need to have (a) X (and Z) are individually related to Y (b) X is related to Z and (c) when X and Z are both used to predict Y then Z is significant but X becomes non-significant (full mediation) or less significant (partial mediation). Our first two logistic models show that both urban setting and and pollution are risk factors for asthma which is consistent with the first requriement. In our logistic model including both variables pollution is highly significant but urban setting has become insignificant which is consistent with (c); specifically with full mediation. The one thing we haven't shown is (b), namely that pollution levels differ between urban and non-urban settings. We could do this using any of a t-test, an ANOVA (with two groups–urban and non-urban) or a simple linear regression (with the urban indicator as the predictor and the pollution level as the outcome).

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.

Solution: Our outcome of interest is having asthma. Risk factors are those that are associated with **higher** lilkelihood of getting asthma. Variables with significant positive coefficients in the log odds (logit) model are risk factors. Equivalently we could look at the odds level (logistic) printout for significant variables with odds ratios greater than 1. It appears that higher pollution levels and having a family history of asthma are associated with higher likelihood of getting asthma. Although breastfeeding has a positive coefficient/odds ratio above 1m it is not significant so we can not say it is a risk factor (and indeed it is generally considered to be protective!) In contrast, higher socio-economic status and being female (gender = 1) appear to be protective. Note that after adjusting for the other factors the coefficient of the urban indicator is negative suggesting a protective effect but this is not even close to being significant.

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.

Solution: For a categorical variable the odds ratio compares the odds of the event of interest for people who do and do not have a characteristic of interest. Here the odds ratio of 3.25 for the family history variable means that **all else equal** a child who has a family member with a history of asthma has odds of getting asthma that are 3.25 times as high as a child who does not have relatives who have had asthma. Note that "all else equal" can be thought of here as comparing two children of the same sex with the same breastfeeding history who have the same SES and live in comparable settinsg (urban/rural; same pollution level.) For a continuous variable the odds ratio tells you about the change in likelihood of the event associated with a particular degree of change in the predictor. Here the odds ratio of .92 for SES means that **all else equal** for every additional point of SES the odds of a child getting asthma go down by 8%.

To calculate the missing regression coefficient for gender we note that the odds ratios are just the exponentiated coefficients or equivalently the regression coeffients are just the natural logs of the odds ratios. Therefore we just need $b_6 = ln(.335) = -1.09$.

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.
Solution: The predicted probability is given by

$$\frac{e^{b_0+b_1X_1+\cdots b_6X_6}}{1+e^{b_0+b_1X_1+\cdots b_6X_6}}$$

The easiest approach is to do the regression plug in to get the log odds and then exponentiate and do the fraction. The child in question lives in a city so $X_1 = 1$. The pollution level is $X_2 = 35$ since pollution was given in thousands and the SES is given as $X_3 = 50$. The child was breastfed for size months so $X_4 = 6$, had parents with asthma so $X_5 = 1$ and is a boy so $X_6 = 0$. Our log odds are therefore given by

$$-.047 + -.077 + .077(35) - .082(50) + .0057(6) + 1.18(1) - 1.09(0) = -.3148$$

Don't forget the intercept! Note also that I deliberately made gender be 0 so that it didn't matter whether you could complete part (d). I'm nice that way :)

Now to get the predicted probability that the boy gets asthma we just do

$$\frac{e^{.3148}}{1 + e^{.3148}} = .42$$

The child has approximately a 42% chance of getting asthma.

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.

Solution: The only difference between the two children is the gender and the pollution level. According to the final model from Problem 1 the pollution level in Los Seraphim in 2011 should be

$$\hat{Y} = 72.07 - 1.195(36) = 29.05$$

Thus the pollution levels should have done down 6 units. For a continuous variable to get the odds ratio associated with a change of Δ we simply raise the odds ratio for a one unit increase to the power Δ. Here $\Delta = -6$ so the odds ratio for the 6 unit drop in pollution is $1.08^(-6) = 0.63$. From the printout the odds ratio for being a girl as opposed to a boy is .33. The girl should be better off on both counts. To get the overall odds ratio we multiply the individual odds ratios and get .21. The girl's odds of getting asthma are 79% lower or only about a fifth as large as her brother's!

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.

Solution: The overall F test tells us whether there is evidence of significant differences among the groups. The classical ANOVA hypotheses for this test are

$H_0 : \mu_W = \mu_{Sp} = \mu_{Su} = \mu_F$–the average symptom level is the same in all four seasons.
H_A : Not all the μ's are equal. There are some differences in average symptom levels by season.

Note that we do NOT set the means to 0! We are not trying to say whether or not there are asthma symptoms in any of the seasons–we are just checking whether the asthma levels, whatver they are, are the same for all seasons. Since the p-value for the F test is .2, much larger than our usual significance level of $\alpha = .05$, we fail to reject the null hypothesis. On the basis of these data we can not say that there are seasonal variations in asthma symptoms (which is a bit surprising....but wait for the rest of the problem!)

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.

Solution: When we fit an ANOVA as a regression the intercept is the mean for our reference group and the coefficients of the indicators for the other groups are simply the differences in means between those groups and the reference group. Here our reference group is Fall. Thus $b_0 = \mu_F = 24.1$. We also need the differences between each of the other seasons and fall. For winter the difference is 5.9, for spring is 13.4 and for summer it is 4.2. The resulting regression equation is

$$\hat{Y} = 24.1 + 5.9X_W + 13.4X_{Sp} + 4.2X_{Su}$$

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.

Solution: None of the comparisons would survive a Bonferroni correction. Even unadjusted only the spring vs fall comparison has a p-value below $\alpha = .05$. With 4 groups (seasons) there are 6 possible pairwise comparisons so we would need to use $\alpha^* = .05/6 = .0083$. The p-value of .039 for fall vs spring is not significant at this new standard. In part (a) we found no evidence of seasonal differences in allergy symptoms so it is not surprising that none of the individual differences came up significant.

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?

Solution: Previously the coefficients for the indicators for winter, spring and summer were all positive. Now the coefficients for winter and summer are negative and both spring and summer seem to differ significantly from fall. The magnitude of the spring coefficient has also decreased substantially. These dramatic changes suggest that at least one of the variables we have added is confounded with season which makes perfect sense. Temperature and humidity certainly are related to season. We would also expect allergies to be worse in spring when pollen levels surge. In fact air pressure and pollution levels could also have seasonal differences.

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.

Solution: The R^2_{adj} value for this model is over 97.5% meaning that these variables explain nearly all the variability in asthma symptoms. In this respect the model is doing a great job. To check how good the predictions are we look at the RMSE compared to the Y values we are trying to predict. Our RMSE = 3.56. From the table at the start of the problem the overall mean symptom level is 30 so our predictions are typically off by a bit over 10%. This is OK but not absolutely fantastic. It seems predicting the symptoms very precisely is hard.

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.)

Solution: We need to perform a partial F test comparing the new model to the original model without the itneractions. The hypotheses for the test are

$H_0 : \beta_9 = \beta_10 = 0$–the interactions between allergy symptoms and humidity and pollution are not significant. The relatiosnhip between asthmas symptoms and whether a child has allergies does not depend on the humidity or pollution levles.
$H_A : \beta_9 \neq 0$ or $\beta_10 \neq 0$ or both–the relationship between allergies and asthma symptoms depends on at least one of humidity and pollution.

The statistic for the partial F test is

$$F_{obs} = \frac{(SSR_{full} - SSR_{red})/m}{SSE_{full}/n - p - m - 1}$$

where p is the number of predictors in the original model, m is the number of variables added in the bigger model. Here we have m = 2 for the two added interactions and p = 8 for the number of predictors in the original model. We get the assorted sums of squares from the two regression printouts. Our test statisti is

$$F = \frac{(51232.8 - 50667.67)/2}{590.9/89} = 42.55$$

I haven't given you an F table. However by now you should now that an F value like this is extremely big and so we will be rejecting the null hypothesis and concluding that the interactions improve the model. The effect of allergies on asthma symptoms depends on at least one of the pollution and humidity levels.

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?
Solution: There are 5 variables involved in this combination: the main effects of humidity, pollution and allergies and the two interaction terms, The easiest way to think about the overall

relationship is to look separately at children who do and don't have allergies. For a child without allerigies the only contributions come from the main effects of humidity and pollution. Since the main effect of humidity is insignificant it seems for children without allergies there is no effect of humidity; however the main effect for pollution is significant (and positive) so we see that for children without allergies asthma symptoms still increase with inc reasing pollution. Now consider children with allergies. The main effect of allergies is positive. This means that if there is no pollution or humidity, children with allergies will have higher asthma symptoms than those who do not. In addition the two interaction terms with allergies are significant and positive. This means that the effects of allergies get even larger as the pollution and humidity levels get higher. Overall this tells us that (i) for all children asthma symptoms increase with pollution, but the rate of increase is greater for children with allergies and (ii) that for children without allergies humidity is not important but for children with allergies asthma symptoms get worse as humidity increases.

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.

Solution: In backwards stepwise we remove the variable with the worst p-value at each stage BUT it is important to respect the hierarchical princple which says that if we have an interaction term then we shouldn't remove any main effect variables that are part of that interaction. Similarly we shouldn't remove one part of a multipart categorical variable that is significant. Here the only variables that are not significant are the indicator for winter, the main effect for humidity and the pressure variable. We can't remove winter because the seasonal components are clearly significant (see the p-values for Spring and Summer) and we can't remove humidity because the allergy by humidity interaction is significant. Therefore we should start by removing the pressure variable. Note however that if we just blindly let a computer package perform backwards stepwise it would remove the humidity variable first on the grounds of it having the highest p-value.

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.

Solution: No one actually got this right on the exam on which it was given so I am saving it for a future year....