2011 Final Exam With Solutions

THE STORY BEHIND THE EXAM: Dr. Caroline R. Ash, a community health scientist and public health policy expert at our favorite school, the University of Calculationally Literate Adults, studies traffic accidents. She is interested in what factors "drive" both non-fatal and fatal accidents and also in the effectiveness of different programs for preventing them. During the exam you will help her analyze data she has collected from several neighboring cities (Los Seraphim, Hollybrick and Beverly Flats) and interpret the results. In particular, Mayor N. D. Burns of Los Seraphim has asked her to comment on the effectiveness of the holiday driving safety program in his city.

Note: Unless otherwise indicated use $\alpha = .05$ for all hypothesis tests on this exam.

Question 1: 'Tis The Season (34 points, 50 minutes)

The winter holiday season is, unfortunately, a peak time for bad traffic accidents. Twenty years ago the city of Los Seraphim instituted an agressive program to promote safe driving practices during this period. The campaign includes advertisements, stepped up traffic patrols, drunk-driving checkpoints and extra penalties for traffic violations. The city has been collecting data on Y, the number of serious traffic accidents during the holidays, each year since the program began. Let X_1 represent Time since the start of the program, so that the intial year was $X_1 = 0$ and the current year is $X_1 = 20$ and let X_2 be the Population of Los Seraphim in each of the same years. Mayor Burns has asked Dr. Ash to look at whether the program has been effective in reducing serious accidents over this period. Dr. Ash has had her graduate student run some preliminary analyses. The results are shown below.

```
. regress NumAccidents Time

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   15.90
       Model |  9443.48679     1  9443.48679           Prob > F      =  0.0008
    Residual |  11285.0679    19  593.950941           R-squared     =  0.4556
-------------+------------------------------           Adj R-squared =  0.4269
       Total |  20728.5547    20  1036.42773           Root MSE      =  24.371

------------------------------------------------------------------------------
NumAccidents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Time |   3.502038   .8782738     3.99   0.001      1.66379    5.340286
       _cons |   137.3211   10.26742    13.37   0.000     115.8312    158.8111
------------------------------------------------------------------------------
*******************************************************************************
. regress NumAccidents Time Population

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  2,    18) =   40.48
       Model |  16957.9351     2  8478.96754           Prob > F      =  0.0000
    Residual |  3770.61959    18  209.478866           R-squared     =  0.8181
-------------+------------------------------           Adj R-squared =  0.7979
       Total |  20728.5547    20  1036.42773           Root MSE      =  14.473

------------------------------------------------------------------------------
NumAccidents |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Time |  -1.924553   1.045449    -1.84   0.082     -4.12096    .2718537
  Population |   .0000983   .0000164     5.99   0.000     .0000638    .0001328
       _cons |   43.24295   16.84962     2.57   0.019     7.843222    78.64268
------------------------------------------------------------------------------
```

Part a (5 points)

Based on these two models do you see evidence of confounding and multicollinearity? Explain briefly and say what the implications are for the interpretation of and conclusions about the regression coefficients in the second model. You may find the summary statistics below useful in justifying your answers.

```
. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
  Population |      4.02    0.248911
        Time |      4.02    0.248911
-------------+----------------------
    Mean VIF |      4.02
```

```
. cor Time Population (obs=21)

             |     Time Popula~n
-------------+------------------
        Time |   1.0000
  Population |   0.8667   1.0000
```

Solution: There is definitely evidence of multicollinearity and confounding in these printouts. Multicollinearity occurs if there is a relationship between two or our of the predictor variables. Here we see that there is a very strong correlation, r = .8667, between the predictors Time and Population. Moreover, the variance inflation factors (VIFs) for Time and Population in the second model are above 4 which is our standard cutoff for indicating that multicollinearity has had a significant impact on the model parameters. Confounding occurs when there is a substantial difference in your interpretation of the relationship between a predictor X and an outcome Y depending on whether or not a third variable Z (the potential confounder) is included in the model. Here Population is a confounder for the relationship between Time and Number of Accidents. In the SLR there is a significant positive relationship between Time and Accidents. However, after adding Population to the model the relationship between Time and Number of Accidents has become negative and in fact is not statistically sigificant if we use a two-sided significance level of $\alpha = .05$. The fact that we have confounding and substantial multicollinearity in this model means that that estimates of the regression coefficients in the second model are not very stable. As indicated by the VIFs their standard errors are inflated meaning there is a great deal of uncertainty about their magnitude and possibly about their significance/sign as well. Thus we need to be cautious about making precise statements about those effect. Note that we are probably much more confident about the sign and significance of the population effect since it's p-value is miniscule but even there we do not have great precision about the magnitude of the effect.

Part b (4 points)

Mayor Burns doesn't like the first printout because he thinks the positive relationship between Time and Number of Accidents implies his safety program hasn't worked. He likes the second printout much better and says that the negative sign on the Time variable in this model supports his theory that the program has reduced the number of accidents. Is there merit to either of the mayor's interpretations? Discuss briefly.

Solution: Mayor Burns obviously didn't major in statistics. There is no merit to his first statement and his second one is also questionable although it is at least not flatly inconsistent with the data. With regard to the first statement, just because the number of accidents went up over time doesn't mean the safety program didn't work. The best measure of how well the safety program worked is probably the accident **rate**, not the number of accidents and it is entirely possible for the number of accidents to go up even while the rate stays constant or even goes down. Indeed if the population increased over time but the rate stayed constant then the number of accidents would rise. As we saw in the second model once we account for population there is no longer a positive relationship between time and accident, which suggests that the reason for the increase in number of accidents was at least partly population growth. (To confirm this it would be helpful to have seen the correlation between Time and Population.) Secondly, even if population growth hadn't accounted for the increase in number of accidents that wouldn't mean the program was at fault. It could be that the number of accidents would have gone up even faster if the program hadn't been in place. The Mayor's second statement is slightly better. The multiple regression indicates that after adjusting for Population our best estimate is that the number of accidents has gone down over time which would be consistent with the program working. We could even claim that this result is significant if we did a one-sided test (which could be justified if we were trying to prove the theory that the program worked) although this is pretty borderline and the instability caused by the multicollinearity/confounding lowers our confidence about the results. However even if we could say with certainty that the accident rate had declined over time that wouldn't be proof that the safety program was the reason even though the two things coincided. Correlation is not causation!

Based on parts (a) and (b) the graduate student decides to create a new outcome variable, PerCapAccidents, which is the number of accidents during the holidays per ten thousand people. The simple linear regression of the new "per capita" accident rate on time is shown below along with corresponding diagnostic plots. Use this output to answer parts (c)-(e).

```
. regress PerCapAccidents Time

      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   19.09
       Model |  .424191401     1  .424191401           Prob > F      =  0.0003
    Residual |  .422297866    19  .022226203           R-squared     =  0.5011
-------------+------------------------------           Adj R-squared =  0.4749
       Total |  .846489267    20  .042324463           Root MSE      =  .14908

------------------------------------------------------------------------------
PerCapAcci~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Time |  -.0234712   .0053726    -4.37   0.000    -.0347163   -.0122262
       _cons |   1.411111   .0628085    22.47   0.000     1.279651    1.542571
------------------------------------------------------------------------------
```

Part c (6 points)

Use the plots from the previous page to discuss whether each of our main regression assumptions is violated. You should make it clear which plot(s) you are using to assess each assumption.

Solution: We assume that the errors are mean 0, independent, constant variance and normally distributed. The first three assumptions are checked with the residual or scatterplot and the normality assumption is checked with the histogram or qq plot. For mean 0 we need the errors to be centered around the 0 line in the residual plot **for all values of X**. In this case the errors are mostly positive for small X, them mostly negative for medium values of X and then positive for large values of X so this assumption is violated. (Note that it is not enough for the positive and negative errors to be balanced overall–they have to be balanced at each X.) The pattern we observed looking at the mean 0 assumption also tells us the independence asusmption is violated. There is a systematic curved shape to the erros that suggests we have fit the wrong model–a curved relationship between accident rate and time would probably fit these data better. Note that the mean 0 and independence assumptions usually go together. For the constant variance assumption we need the spread of the band of the residuals to be the same width for all X. Note that this doesn't mean that the spread has to go equally high above and below the 0 line. If the mean 0 and independence asusmptions are violated the band will be curved but it can still be the same "fatness" all the way along. Here we don't have many points at each value of X so it is hard to judge the width of the band well. Some people said it was wider in the middle and narrower at the ends while others said it was OK. We accepted either answer as long as it was clear you knew what you were looking for.

Part d (5 points)

Do you see any possible outliers or influential points? Explain (i) based on the scatter plot and (ii) based on the diagnostic statistics below. Do you think any of the points are problematic enough that they need to be removed? Discuss briefly.

Time	StudentizedRes	Leverage	CooksDistance	DFBeta(Time)
0	6.29	.18	1.41	-2.49
11	-1.17	.05	.03	-0.04
21	1.02	.18	.11	0.41
Next Highest	-1.08	.15	.07	0.31

Solution: First, looking at the scatterplot, we see that the Time=0 point looks quite separated from the others. This point looks like an outlier in Y. This point is also the worst based on the diagnostics shown above. It's studentized residual is a whopping 6.29, well above our usual threshold of 2 or 3. None of the other points comes close to being a problem in this respect. It is also tied for the highest leverage value at .18. Our threshold for being high leverage is $2(p+1)/n$ where p is the number of predictors and n is the sample size. Here $p = 1$ and $n = 21$ so the cutoff would be $2(1+1)/21 = .19$ which our point doesn't meet although it is close. None of these points has super high leverage although naturally the points at the edge of the data set (Time = 0 and Time = 20) are the highest. The Time=0 point also has by far the largest Cook's Distance and DFBeta values suggesting it is the most influential. Our rough rules of thumb said that points with

Cook's Distances above 1 or DFBetas above 2 are influential. Thus according to the diagnostics our point is an influential outlier and may well be affecting the fit of the line. However you can't just say on this basis that the point should be removed. To justify that we'd need some reason for saying the point was a mistake or not representative of the population of interest. If that isn't the case then you'd need to fit the model with and without the point and report exactly how big an effect it is having on your conclusions (e.g. significance and direction of the relationship.) However here we have a different problem: The point is probably looking like an outlier because we have fit the wrong model. The error diagnostics suggested that a curvilinear model would fit the data better. If we sketch in a curved shape for the relationship suddenly the point at Time=0 is right in line with the others. There is one semi-reasonable argument you could make for removing this point. Since it represents the first year the program was implemented you could speculate that the safety rules had not had time to fully take effect or that there were still some bugs in implementation and thus the period starting with Time=1 would better reflect the effects of the program. However it is generally a good idea to look at change from the baseline to understand the progression over time.

Part e (4 points)

Mayor Burns like this model too. He decides to predict what the per capita accident rate will be like 50 years from now. Does the answer he gets make real-world sense? If so why? If not what has gone wrong?

Solution: We note that the program started 20 years ago so 50 years from now will correspond to Time= 70. Plugging in this value our predicted accident rate is

$$\hat{Y} = b_0 + b_1 Time = 1.41 - .0235(70) = -.235$$

or negative .235 accidents per 10,000 people or about negative 2 accidents per 100,000 peopple. Obviously this doesn't make any sense–you can't have negative accidents! What has gone wrong of course is that we have extrapolated WAY outside the range of our data. It is highly unlikely that the model that works for accident rate now will still be correct in 50 years when the city may have changed completely. Moreover we know from parts (c) and (d) that the model we have fit is not even the right one for the current data so if we extrapolate this far we guaranteed to be in trouble. Mayor Burns is offering even more evidence that he knows nothing about statistics! Note that some people plugged in Time=50 rather than Time=70. If you do that you get a predicted accident rate of around .235. While this number is physically possible it doesn't fit at all with the scatterplot which shows the accident rate leveling off at about 1 per 10,000 people so you can still tell that the extrapolation is unreasonable.

Next the graduate student decides to try out some transformations of the time variable. Specifically she tries (i) adding a quadratic term, TimeSq, (ii) using LogTime and (iii) fitting an inverse model in which the predictor is InvTime=1/(1+Time). The corresponding printouts are shown below.

```
. regress PerCapAccidents Time Timesq
      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  2,    18) =   28.05
       Model |  .640837754     2  .320418877           Prob > F      =  0.0000
    Residual |  .205651513    18  .011425084           R-squared     =  0.7571
-------------+------------------------------           Adj R-squared =  0.7301
       Total |  .846489267    20  .042324463           Root MSE      =  .10689
------------------------------------------------------------------------------
PerCapAcci~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Time |  -.0856247   .0147838    -5.79   0.000    -.1166842   -.0545651
      Timesq |   .0031077   .0007137     4.35   0.000     .0016083     .004607
       _cons |    1.60793   .0638021    25.20   0.000     1.473887    1.741974
------------------------------------------------------------------------------
*******************************************************************************
. regress PerCapAccidents LogTime
      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =   73.14
       Model |   .67194482     1   .67194482           Prob > F      =  0.0000
    Residual |  .174544447    19   .00918655           R-squared     =  0.7938
-------------+------------------------------           Adj R-squared =  0.7829
       Total |  .846489267    20  .042324463           Root MSE      =  .09585
------------------------------------------------------------------------------
PerCapAcci~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     LogTime |  -.5162129   .0603585    -8.55   0.000    -.6425447   -.3898812
       _cons |   1.660861    .060384    27.51   0.000     1.534476    1.787246
------------------------------------------------------------------------------
*******************************************************************************
. regress PerCapAccidents InvTime
      Source |       SS       df       MS              Number of obs =      21
-------------+------------------------------           F(  1,    19) =  212.33
       Model |  .776962532     1  .776962532           Prob > F      =  0.0000
    Residual |  .069526735    19  .003659302           R-squared     =  0.9179
-------------+------------------------------           Adj R-squared =  0.9135
       Total |  .846489267    20  .042324463           Root MSE      =  .06049
------------------------------------------------------------------------------
PerCapAcci~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     InvTime |   .8970007    .061559    14.57   0.000     .7681561    1.025845
       _cons |    1.02069   .0169836    60.10   0.000     .9851429    1.056237
------------------------------------------------------------------------------
```

Part f (4 points)

Give a brief interpretation of the coefficient of the LogTime variable and in particular explain what the negative sign tells you in real-world terms. How does this fit with the mayor's theory?

Solution: The literal interpretation of the value $b_1 = -.516$ is that for every one unit increase in LogTime the accident rate **drops** by about half an accident per 10,000 people. We need to be a bit more precise though. First, what does a one unit change in LogTime mean? LogTime goes up by 1 when Time on the original scale is **doubled** This is assuming that we are using Log Base 2 which I didn't actually say. (Thus we get a drop of .5 in the accident rate going from time 1-2, 2-4, 4-8, 8-16, etc.) If it were log base 10 or natural log it would correspond to time being **multipled** by a factor of 10 or 2.73 respectively. The time it takes for each successive drop of .5 per 10,000 people in the accident rate gets longer and longer. What this means is that the accident rate is declining over time but at an ever slower rate–the log function decreases less rapidly than a straight line. However the log function does NOT level off–our regression equation will get arbitrarily large and negative if we extrapolate far enough. Since the accident rate is decreasing over time this model is **consistent** with the theory that the program is working although as we have noted before we can not prove the program is **causing** the decrease–correlation is not causation.

Part g (6 points)

Rank the three transformations from best to worst. Briefly explain how you are making your choice and say whether the best model makes real-world sense.

Solution: We can rank the models based on R^2_{adj} (for which higher values are better) or the RMSE (for which lower values are better). Using these criteria the InverseTime model is the best, followed by the LogTime model and the worst is the TimeSquared model (though even it is fairly good.) Note that it is not a good idea to use raw R^2 which always gets bigger as you add more variables, nor can we directly compare the F values and their p-values since the models do not all have the same number of variables and what counts as a significant F test value depends on the degrees of freedom. Some people said we couldn't compare RMSE's because the transformations meant the models weren't on the same scale. This would be correct if we'd transformed Y but our transformations were all on the predictor, X, which doesn't effect the scale. RMSE is un units of Y and here the Y variable was always the same.

The inverse model $\hat{Y} = 1.02 + .897/(1 + Time)$ does make a good deal of practical sense here. As Time goes by, $1/(1 + Time)$ gets smaller and smaller so the equation levels off to a value of 1.02. Since the sign on the inverse term is positive the function is actually decreasing in Time. (To see this note that at Time = 0 our predicted value is 1.917 and that the amount we add to the b_0 value of 1.02 gets smaller with each step.) This suggests that initially the program has a fairly large impact on the accident rate but that there are diminishing returns over time until eventually everyone has learned or adjusted to the program conditions and there is no further change. Of course there will still be some accidents because there are still effects of weather, bad roads, bad luck and people who simply ignore the program which is why the limiting accident rate isn't 0.

Question 2: Intercity Interactions (29 points, 40 minutes)

Dr. Ash points out to her graduate student that there would be more convincing evidence for the effectiveness of the Los Seraphim traffic safety program if the time trend found at the end of Question 1 could be compared with those in some other similar cities. Data are available from the neighboring communities of Hollybrick (which has no accident prevention program) and Beverly Flats (which has been running ads promoting safe driving during the holidays but has no other special programs). The graduate student creates indicator variables for each city. LS = 1 if the city is Los Seraphim and 0 otherwise, HB=1 if the city is Hollybrick and 0 otherwise and BF=1 if the city is Beverly Flats and 0 otherwise. The printout for a regression of per capita accident rate on the indicators HB and BF is shown below.

```
. regress PerCapAccidents HB BF
      Source |       SS       df       MS              Number of obs =      63
-------------+------------------------------           F(  2,    60) =   14.26
       Model |  1.98425675     2  .992128374           Prob > F      =  0.0000
    Residual |  4.17565874    60  .069594312           R-squared     =  0.3221
-------------+------------------------------           Adj R-squared =  0.2995
       Total |  6.15991548    62  .099353476           Root MSE      =  .26381
------------------------------------------------------------------------------
PerCapAcci~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HB |   .3874778   .0814127     4.76   0.000     .2246282    .5503275
          BF |   .3644091   .0814127     4.48   0.000     .2015595    .5272588
       _cons |   1.176399   .0575675    20.44   0.000     1.061247    1.291551
------------------------------------------------------------------------------
```

Part a (5 points)

Based on this model does there seem to be **overall** evidence of a significant difference in accident rates among the three cities? State the matehmatical hypotheses you would be testing using (i) the regression framework and (ii) the classical ANOVA framework and your real-world conclusions.

Solution: To see whether there is overall evidence of a difference in accident rates we need to do an F test. In the regression framework our hypotheses are:

$H_0 : \beta_{HB} = \beta_{BF} = 0$–the mean accident rates for HollyBrick and Beverly Flats are no different from that for Los Seraphim.
H_A : At least one of $\beta_{HB}, \beta_{BF} \neq 0$–there are some differences among the cities.

In the classical ANOVA framework we compare the means of each group directly rather than looking at differences from a reference group. Here we would have

$H_0 : \mu_{LS} = \mu_{HB} = \mu_{BF}$–the cities all have the same average accident rate.
H_A : the means are not all equal; the accident rates differ across cities.

Note that in the first set of hypotheses I am testing whether the parameters are equal to 0–I am checking for a non-zero **difference** between the groups. In the second set of hypotheses I do NOT need the "=0" piece because I am looking at the actual accident rates and the goal is to see whether they are the same, not whether they are equal to 0-.

From the regression printout the F statistic for the test is $F_{obs} = 14.26$ and the corresponding p-value is essentially 0 so we reject the null hypothesis and conclude that there is evidence of a difference in accident rates across the cities.

Part b (5 points)

Use the printout to find the estimated mean accident rate for each of the three cities and from that obtain the regression equation you would have gotten if Hollybrick had been used as the reference city.

Solution: In the regression ANOVA framework, the intercept gives the mean value of Y for the reference group. Here Los Seraphim is the reference group so we see that the average accident rate there is 1.176 accidents per 10,000 people. To get the means for the other cities we have to plug in their respective indicators:

HollyBrick: $\hat{Y} = 1.176 + .387(1) + .364(0) = 1.563$
Beverly Flats: $\hat{Y} = 1.176 + .387(0) + .364(1) = 1.54$

Now to get the equation with HollyBrick as the reference city we simply use it's mean as the intercept and for the coefficients of the other two indicators we use their difference from the reference group. We note that LS - HB = 1.176 - 1.563 = -.387 (this is just the negative of what we got as the coefficient of HollyBrick when Los Seraphim was the reference.) Similarly for Beverly Flats we get BF-HB = 1.54 - 1.563 = -.023. The resulting regression equation is

$$\hat{Y} = 1.563 - .387LS - .23BF$$

Part c (6 points)

Find the effect sizes associated with the differences in accident rates between the three cities and say whether they are large by the standards we learned in class. Do you think this study **turned out** to have been adequately powered? Would you have said it was adequately powered **before it started?** Explain briefly. The following printout may be helpful.

```
. sampsi 0 .89, sd1(1) sd2(1) n1(20) n2(20) alpha(.05)

Estimated power for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:
```

```
         alpha =   0.0500  (two-sided)
            m1 =        0
            m2 =      .89
           sd1 =        1
           sd2 =        1
sample size n1 =       20
            n2 =       20
          n2/n1 =     1.00
Estimated power:
         power =   0.8036
```

Solution: The effect sizes here are just the difference in means between the cities divided by the within group standard deviation. SInce this is an ANOVA we are assuming all the groups have the same standard deviation and it is estimated by RMSE = .264 from the regression printout. Our effect sizes are therefore:

LS vs HB: $d = (1.563 - 1.176)/.264 = 1.47$
LS vs BF: $d = (1.54 - 1.176)/.264 = 1.38$
HB vs BF: $d = (1.563 - 1.54)/.264 = .087$

The first two effects, comparing Los Serpahim to each of the ohter cities, are very large (in fact our rough rule of thumb says anything over about .8 of a standard deviation is a large effect) while the difference between HollyBrick and Beverly Flats is very small. **After the fact** it seems that the study was well powered. The first two observed effects were very large and the tests came out significant. Any time you get a significant result it means you were able to detect the result you were looking for. We didn't get a significant result for HB vs BF but it looks like there really just isn't a difference there. However **before the study** it would **not** have looked like the study was well powered. From the STATA printout our minimum detectable effect with 80% power is d = .89 which is a large effect. This means that our study was only likely to see effects if they were rather large. Unless we had some argument for why we would expect the effects to be large this would not be considered adequate power.

Part d (Optional Bonus:)

Given your findings in Question 1, an ANOVA is probably not an appropriate model for these data. Give two reasons, one having to do whether whether the model will fit well and one having to do with whether the data fulfill the model assumptions.

Solution: From Question 1 we know that there is a significant time trend in the accident rates, at least in Los Seraphim. Since we are leaving out a highly significant predictor it seems unlikely our model will describe the accident rates particularly well. Moreover, the fact that there is a time trend means that if we ignore it our regression assumptions will be violated. In particular the residuals (which here are differences from the city mean) will not be independent (they will be correlated over time) and they will also not be normally distributed. The latter can be seen by noting that for Los Seraphim at least the accident rate started out high and then leveled off so that

for the last 10 years most of the Y values were very similar whle the ones for the first few years were much higher. This will lead to a skewed distribution.

After listening to Dr. Ash explain the answer to part (d), the graduate student fits a new model with the two city indicators, inverse time and the interactions between city and inverse time. The new printout is shown below. Use it to answer the remaining questions.

```
. regress PerCapAccidents HB BF InvTime InvTimeHB InvTimeBF

      Source |       SS       df       MS              Number of obs =      63
-------------+------------------------------           F(  5,    57) =   57.25
       Model |  5.13697321     5  1.02739464           Prob > F      =  0.0000
    Residual |  1.02294228    57  .017946356           R-squared     =  0.8339
-------------+------------------------------           Adj R-squared =  0.8194
       Total |  6.15991548    62  .099353476           Root MSE      =  .13396

------------------------------------------------------------------------------
PerCapAcci~s |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HB |   .5416091   .0531904    10.18   0.000     .4350972     .648121
          BF |   .2478435   .0531904     4.66   0.000     .1413316    .3543554
     InvTime |   .8970007   .1363267     6.58   0.000     .6240112     1.16999
   InvTimeHB |  -.8879118    .192795    -4.61   0.000    -1.273977   -.5018464
   InvTimeBF |   .6715053    .192795     3.48   0.001     .2854398    1.057571
       _cons |    1.02069   .0376113    27.14   0.000     .9453747    1.096005
------------------------------------------------------------------------------
```

Part e (7 points)

Did adding the time and interaction terms improve the model relative to the one with just the city indicators? Write down the null and alternative hypotheses corresponding to this test both mathematically and in words, obtain the test statistic and say whether or not you think the test will be significant (you can't get the exact p-value but you should be able to make an educated guess!) Explain briefly how you could get STATA or SAS to perform this test for you.

Solution: Here we are being asked to compare the model from part (a) to the new model using a partial F test. Our hypotheses are

$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$–none of the time or interaction terms improves the model.
H_A : At least one of the $\beta's \neq 0$–the second model is an improvement over the first.

Our test statistic is

$$F = \frac{(SSR_{full} - SSR_{red})/m}{SSE_{full}/(n-p-m-1)} = \frac{(5.14 - 1.98)/3}{1.02/57} = 58.86$$

Note that $m = 3$ is the difference in the number of variables between the two models, $n = 63$ is the sample size and $p = 2$ is the number of predictors in the original model. To get the exact p-value we would need STATA or SAS. However experience tells us that F values even around 5 or 6 are often significant (depending on the exact degrees of freedom) and a value like 58.86 is huge. Moreover all of the time and interaction terms are individually highly significant in the second model. All of these things together suggest that we will be rejecting our null hypothesis. The second model is definitely a better fit for the data than the first.

There are several ways we could get STATA or SAS to perform this test for us. One would be to use a follow-up test or contrast statement of the form "test InvTime = InvTimeBF = InvTimeHB = 0". Another would be to treat city as a multilevel categorical variable so that the output would automatically give the p-values for whether city and it's interaction with time were significant. Note though that this latter approach isn't really quite good enough since we are adding not just the interaction but also the time variable. If you got confused from the wording of the problem about whether you were just testing for the interaction or for all three variables we were pretty merciful on the scoring.

Part f (6 points)

Give an intuitive explanation of what an interaction means in the context of this model. Then describe as carefully you can the differences in the trends of the accident rates over time for the three cities. What does this suggest about whether the safety programs have been helping to reduce accident rates? (Note: As always with interactions you may find it helpful to plug in the values of the indicators and get the resulting equations for each city.)

Solution: The general definition of an interaction between two predictors X_1 and X_2 is that the relationship between Y and X_1 depends on the **value** of X_2 and vice versa. Here that means that the relationship between accident rate and time depends on which city you are in (i.e. the cities have different time trends for accident rate) or equivalently that the differences between the cities change over time. If we plug in the indicators for the cities we get the following:

LS: $\hat{Y} = 1.02 + .897InvTime$
HB: $\hat{Y} = 1.02 + .542 + .897InvTime - .888InvTime = 1.562 - .009InvTime$
BF: $\hat{Y} = 1.02 + .248 + .897InvTime + .672InvTime = 1.268 + 1.569InvTime$

What we see is that in Hollybrick the accident rate is almost flat at 1.562 accidents per 10,000 people. In both Los Seraphim and Beverly Flats the accident rate is decreasing over time (which makes sense since they both instituded safety programs at the start of the period covered by the data). However the accidents are decreasing at different speeds and have different limiting rates. In particular we eventually expect Hollybrick to have the highest accident rate, around 1.562, followed by Beverly Flats at 1.268 and Los Serpahim will have the loest rate around 1.02 accidents per 10,000 people. This all makes perfect sense since Hollybrick has no safety program (so we wouldn't expect its accident rate would change), Berverly Flats has a modest one and Los Seraphim has the most agressive program. Our evidence that the programs work is certainly stronger now in that we have data from three otherwise probably quite comparable cities whose accident rates correspond to what we would expect given their respective safety programs. However since this is not a

controlled experiment and we have not explicitly accounted for other factors which might differ between the cities and be related to the outcome we certainly don't have definitive proof that it is the programs that have caused the differences we see.

Part g (Optional Bonus)

Note that the R^2 value for this model is much lower than that in the final model of Question 1. Is this surprising? Explain briefly.

Solution: This isn't very surprising at all as the models are for two different data sets. The first model covers only Los Seraphim while the second model adds two additional cities. It is entirely possible that there is more variability in accident rates in HollyBrick and Beverly Flats or that the inverse time model we selected based on Los Seraphim doesn't fit the other cities so well. Either of these things would lead to the overall R^2 value for the second model being lower.

Question 3: Don't Drink and Derive (37 points, 50 minutes)

Next, rather than looking at trends over time, Dr. Ash decides to focus on the factors associated with traffic fatalities using data from the last 300 serious traffic accidents in Los Seraphim. She has recorded whether or not the accidents involved fatalities ($Y = 1$ for yes and $Y = 0$ for no) as well as the predictors listed below. Note that all personal characteristics (e.g. gender, drinking status) refer to the at fault driver or vehicle.

X_1: Holiday season. $X_1 = 1$ if the accident took place during the holidays and 0 if not.
X_2: Drinking status. $X_2 = 1$ if the driver was legally intoxicated and 0 if not.
X_3: Gender. $X_3 = 1$ for female and 0 for male.
X_4: Age, in years.
X_5: Age squared.
X_6: Speed, in miles per hour.
X_7: Dark. $X_7 = 1$ if the accident took place after sunset and 0 otherwise.
X_8: Speed by Dark interaction. $X_8 = X_6 * X_7$.
X_9, X_{10}: Indicators for the number of cars involved in the accident. This is divided into three categories: solo (1 car), pair (2 cars) or multi (3 or more cars). $X_9 = 1$ if it was a "solo" accident and 0 otherwise and $X_{10} = 1$ if it was a "multi-car" (3 or more vehicles) accident.

Dr. Ash suspects that accidents are more likely to be fatal during the holidays than at other times. To test her theory she performs a simple logistic regression with X_1 as her predictor.

```
. logistic fatal holiday
Logistic regression                                     Number of obs   =        300
                                                        LR chi2(1)      =       3.92
                                                        Prob > chi2     =     0.0478
Log likelihood = -161.02363                             Pseudo R2       =     0.0120
------------------------------------------------------------------------------
       fatal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     holiday |    1.77907   .5120566     2.00   0.045     1.012041    3.127433
------------------------------------------------------------------------------
```

Part a (4 points)

Write down the mathematical hypotheses for the test Dr. Ash would use to try to prove her theory and give the corresponding p-value and your real-world conclusions.
Solution: Dr. Ash is trying to prove the likelihood of a fatal accident is **higher** during the holidays which is a 1-sided test. Her hypotheses are therefore

$H_0 : \beta_1 \leq 0$–The likelihood of a car accident being fatal during the hiliday season is the same or lower as during the non-holiday period.
$H_0 : \beta_1 > 0$–The likelihood of an accident being fatal is higher during the holidays.

To get the 1-sided p-value we have to divide STATA's p-value by 2 so we get .0225. Since this is less than our significancel level of $\alpha = .05$ we reject the null hypothesis and conclude that the likelihood of an accident being fatal is higher during the holidays.

Part b (Optional Bonus)

The table below shows the actual numbers of accidents (fatal and non-fatal) for each of the time periods considered (holiday or not). Show how you could calculate the odds ratio comparing holiday to non-holiday periods based on the table and verify that your answer is consistent with the logistic regression printout.

	Fatal		
Holiday	Yes	No	Total
Yes	27	60	87
No	43	170	213
Total	70	230	300

Solution: The odds ratio we want simply compares the odds of an accident being fatal during the holidays to the odds of an accident being fatal during the non-holiday period. The odds of an event, F, are given by $p_F/(1-p_F)$. We can get the estimated probability of a fatality during each period easil from the table. For the holiday period the odds are $(27/87)/(60/87) = 27/60 = .45$. For the non-holiday period the odds are $(43/213)/(170/213) = 43/170 = .252$. The resulting odds ratio is OR $= .45/.253 = 1.779$ which is exactly the value on the logistic regression printout.

Part c (6 points)

Dr. Ash believes that the main reason the traffic fatality rate goes up during the holidays is that people are more likely to drink and drive at this time of year. To test this theory she has fit several additional models which are shown on the next page. Explain (i) Why Dr. Ash's theory is one of mediation (ii) how the models she has fit correspond to the steps necessary for testing mediation and (iii) whether she has evidence for either full or partial mediation.

Solution: Mediation occurs when a variable X is related to Y **through** it's effect on an intermediate variable Z. The idea is that X causes Z which in turn causes Y. Here Dr. Ash believes that the likelihood of fatal accidents increases during the holidays because people drink and drive more during this period and drunk driving is dangerous. This is a classic mediation set up. We can not prove mediation using regression models because regression models do not show causality, they simply show whether or not there is an association. However we can use regression techniques to see whether our data are consistent with mediation. Specifically we check whether (a) the predictor, X, is individually related to the outcome Y (if it isn't there is no relationship between X and Y for Z to mediate!) (b) whether the the predictor, X, is related to the mediator Z (if it isn't then the relationship between X and Y can't go through Z!) and (c) whether when X and Z are both included in the model for predicting Y, Z is significant and X either becomes insignificant (full mediation–the entire effect of X is explained by Z) or less significant (partial mediation–some but not all of the effect of X is explained by Z). In question (a) above we showed that the predictor variable, holiday season, was related to the outcome of interest, namely whether or not an accident

was fatal. This is step (a) in our mediation sequence. The models below show that whether or not a person is drunk (our potential mediator) is related to whether or not the accident is fatal (not directly required but often done as part of the mediation test); that season is associated with drunkeness (step (b) of the mediation sequence); and that when season and drunkeness are both included in the model drunkeness is significant (p-value = .007) but season is not (.697). Since season has become completely insignificant these results are associated with complete mediation.

```
. logistic fatal drunk

Logistic regression                                     Number of obs   =        300
                                                        LR chi2(1)      =      10.98
                                                        Prob > chi2     =     0.0009
Log likelihood = -157.48995                             Pseudo R2       =     0.0337

------------------------------------------------------------------------------
       fatal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drunk |   2.538462   .7116213     3.32   0.001     1.465374    4.397367
------------------------------------------------------------------------------
*******************************************************************************
. logistic drunk holiday

Logistic regression                                     Number of obs   =        300
                                                        LR chi2(1)      =     143.18
                                                        Prob > chi2     =     0.0000
Log likelihood = -119.36242                             Pseudo R2       =     0.3749

------------------------------------------------------------------------------
       drunk | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     holiday |   35.92063   12.82827    10.03   0.000      17.8384    72.33227
------------------------------------------------------------------------------
*******************************************************************************
. logistic fatal drunk holiday

Logistic regression                                     Number of obs   =        300
                                                        LR chi2(2)      =      11.14
                                                        Prob > chi2     =     0.0038
Log likelihood = -157.41108                             Pseudo R2       =     0.0342

------------------------------------------------------------------------------
       fatal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drunk |   2.817059   1.081768     2.70   0.007     1.327178     5.97947
     holiday |   .8537866   .3405592    -0.40   0.692     .3906801    1.865853
------------------------------------------------------------------------------
```

Dr. Ash is now ready to see what other factors are associated with traffic fatalities. Based on the findings from part (a)-(c) she decides to fit a model with all the variables except holiday season (i.e. she uses X_2-X_{10}). The resulting printouts are shown below. Use them to answer the remainder of the problem.

```
. logit fatal drunk gender age agesq speed dark speedbydark solo multi
Logistic regression                                 Number of obs   =        300
                                                    LR chi2(9)      =     138.55
                                                    Prob > chi2     =     0.0000
Log likelihood = -93.708786                         Pseudo R2       =     0.4250

------------------------------------------------------------------------------
       fatal |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drunk |   .8964695   .3912408     2.29   0.022     .1296516    1.663287
      gender |  -.9672396   .4247317    -2.28   0.023    -1.799698   -.1347807
         age |  -.1507639   .0577549    -2.61   0.009    -.2639614   -.0375664
       agesq |   .0015265   .0005447     2.80   0.005     .0004589     .002594
       speed |   .0261336   .0288708     0.91   0.365     -.030452    .0827193
        dark |   2.726086   2.314635     1.18   0.239    -1.810514    7.262686
 speedbydark |   .0183128   .0079621     2.30   0.021      .016752     .019873
       solo |   1.178641   .4137755     2.85   0.004     .3676563    1.989626
      multi |   .9519267   .3399738     2.80   0.005     .2855781    1.618275
      _cons |  -3.367394   2.605579    -1.29   0.196    -8.474236    1.739448
------------------------------------------------------------------------------
*******************************************************************************
. logistic fatal drunk gender age agesq dark speedbydark solo multi
Logistic regression                                 Number of obs   =        300
                                                    LR chi2(9)      =     138.55
                                                    Prob > chi2     =     0.0000
Log likelihood = -93.708786                         Pseudo R2       =     0.4250

------------------------------------------------------------------------------
       fatal | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drunk |   2.450935   .9589057     2.29   0.022     1.138432    5.276629
      gender |   .3801309   .1614537    -2.28   0.023     --------    --------
         age |   .8600507   .0496721    -2.61   0.009     .7680032    .9631305
       agesq |   1.001528   .0005455     2.80   0.005     1.000459    1.002597
       speed |   1.026478   .0296352     0.91   0.365      .970007    1.086237
        dark |   15.27299    35.3514     1.18   0.239       .16357    1426.083
 speedbydark |   1.018481   .4218738     2.30   0.021     1.016893    1.020072
       solo |   3.249956   1.344752     2.85   0.004     1.444346    7.312801
      multi |   2.590696   1.021375     2.80   0.005     1.330531    5.044381
------------------------------------------------------------------------------
```

Part d (4 points)

Based on the printouts on the previous page explain briefly what would happen at the first step of an **automated** backwards stepwise model selection procedure. Do you agree with that step? If so, explain why; if not explain what you would do instead.

Solution: An automated backwards stepwise procedure begins by removing the variable with the highest p-value, assuming there is one which is insignificant (p-value $> .05$). Here the worst predictor appears to be speed with a p-value of .365. However we note that speed is part of an interaction trio (speed, dark and a speedbydark interaction) and the hierarchical principal tells us that if the highest order term in a set of grouped variables is significant then we should keep the whole set. Here the speed by time of day interaction is significant (p-value .021) which means we should keep both the speed variable and the time of day indicator even though they aren't individually significant. Note that these are the only two variables in the model that are not significant so in fact we shouldn't remove anything from this model.

Part e (5 points)

Give a brief interpretation of the odds ratio for gender and it's corresponding confidence interval. (Note that the confidence interval has mysteriously disappeared from the printout so you will need to compute it first!)

Solution: From the second part of the printout we see that the odds ratio for gender is .38. This means that a woman (gender=1) has odds of being a fatal accident only .38 times as high as an otherwise equivalent man. Alternatively we can say that, all else equal, a woman has 62% lower odds of being in a fatal accident than a man.

To get the confidence interval for the odds ratio we simply exponentiate the confidence interval for the regression coefficient from the first printout. The resulting interval is $[e^{-1.8}, e^{-.135}] = [.165, .874]$. This means that we are 95% sure that a woman has odds somewhere between 12.6% to 83.5% lower of being in a fatal accident than a man, all else equal. Note that the interpretation of the confidence interval is exactly the same as for the odds ratio itself–it just gives the best case and worst case possibilities.

Part f (4 points)

Is a linear relationship adequate to describe the relationship between age and the log odds of a fatal accident or is a curvilinear relationship superior? Briefly justify your answer using an appropriate p-value and describe in real-world terms what the shape of the relationship is telling you.

Solution: The model includes both linear and quadratic terms in age. Since the age squared term is highly significant (p-value = .005) we conclude that a curvilinear model describes the relationship between age and log odds of a fatal accident better than a simple linear relationship. The fact that the sign on the age squared term is positive means that (on the log odds scale) the shape of the relationship is an up-opening parabola. Specifically for lower ages the likelihood of a fatal accident is higher, then decreases for a while as age increases, but then changes direction and increases

again for higher ages. In practical terms this means that very young and very old drivers are at the highest risk of being in a fatal accident.

Part g (5 points)

Based on the estimates from the printout, rank the three crash types (solo, pair or multi-car) from lowest to highest likelihood of involving a fatality. Do you think that solo and multicar crashes are significantly different? Explain your reasoning and say how you could formally test for this in STATA or SAS using an appropriate follow-up contrast.

Solution: The pair or two-car accident is the reference category in this model. Since the coefficients for the indicators of solo and multicar accidents are both positive it seems both of these sorts of crashes are more likely to involve fatalities. (This makes a certain amount of sense. To have a solo accident usually means you have gone fairly badly out of control. A multicar accident is usually also fairly serious. Two-car accidents on the other hand may be more likely to be simple fender benders....) From the printout it doesn't look like the solo and multicar accidents are very different in terms of their likelihood of being fatal. In particular, their regression coefficients are very similar and the resulting confidence intervals on either the lod odds or odds ratio scales have a huge degree of overlap. To test this formally we could use a follow-up test or contrast statement like "test solo = multi". Note that in this case we do NOT test whether both coefficients are equal to 0–that would be testing whether either of these accident types was different from a two-car accident. All we want to test here is whether the two coefficient are the SAME.

Part h (6 points)

Find the predicted probability of a fatality in a solo crash involving a 20 year old male drunk driver going 80 miles per hour at night.

Solution: All we have to do is plug in to the estimated regression equation and then do the appropriate exponentiation. We start by getting the predicted log odds of a fatal accident:

$$L = -3.37 + .896(1) - .967(0) - .151(20) + .0015(20^2) + .026(80) + 2.726(1) + .018(1)(80) + 1.179(1) + .952(0) = 2.531$$

To solve for the probability given the log odds, L, we just use the formula

$$p = \frac{exp^L}{1 + exp^L} = \frac{exp^{2.531}}{1 + exp^{2.531}} = .926$$

The probability that such an accident is fatal is over 90%!

Part i (3 points)

Suppose you wanted to use a Bonferroni correction to adjust for the number of tests for individual predictors in this model. What significance level would you use and would any of the variables remain significant?

Solution: There are m=9 predictor variables in our model so as our new significance level we would use $\alpha^* = \alpha/m = .05/9 = .0056$. The only variables that remaion significant after this correction are age squared and the accident type indicators, solo and multi.

Part k (Optional Bonus)

Suppose a solo crash had been used as the reference category instead of a two-car crash. Find the corresponding regression coefficients and odds ratios for the new model, briefly explaining your reasoning.

Solution: The current reference category is a two-car accident. From the given regression coefficients we know that the log odds for a solo accident are 1.179 higher than for a two-car accident so the log odds for a two-car accident must be 1.179 lower than for a solo accident. By taking the difference between the coefficients for the multi-car accident and the solo accident we see that the log odds for a multicar accident are .227 lower than for a solo accident. Thus the coefficients for the indicators of "two car accident" and "multicar accident" would be respectively -1.179 and -.227. To get the corresponding odds ratios we exponentiate getting ORs of .31 and .80 respectively. Of course we could have done this on the odds ratio scale first and worked backwards...