Biostatistics 200A, 2010 Midterm With Solutions

General Comments:

- There was a lot of variability in the performance on this exam. Below I give some summary statistics and rough grade ranges. You should take these as estimates, not as the exact score needed to get a particular grade for the class. That cutoff will depend on the difficulty of the final, the performance on the projects, and so on. However these ranges should give you an idea of how I felt about the exam over all. If you are concerned about your performance, please come see me and we can work out a strategy for the rest of the class. Do remember that if you do better on the final than on the midterm then the weight on the midterm gets reduced to 10% so there is still plenty of opportunity to raise your grade!

- Please read the solutions carefully, even for parts on which you got full credit, as I use them as an opportunity to try to give further insight about the material. Because of this my solutions are **much** more detailed than yours would have needed to be. If after reading my answers you are not sure why you lost points, feel free to come ask.

Statistic	Value
Mean	78.86
Median	80
Standard Deviation	12.73
Maximum	94
Minimum	40.5
1st Quartile	74
3rd Quartile	88.75

Approximate Grade	Score Range	Number In Range
A	90+	12
A-	83-89.5	12
B+	75-82.5	14
B	66-74.5	6
B-	55-65.5	4
Below	< 55	3

THE STORY BEHIND THE EXAM: Professor Urtha Green, an environmental health scientist at my favorite college, the University of Calculationally Literate Adults, is studying fine particle pollutants at elementary schools in the city of Los Seraphim. She is particularly interested in differences between indoor and outdoor pollutant levels and in how factors like distance from the freeway, temperature, location, weather and time of day affect the air quality. During the exam you will help her analyze some of her data and interpret the results.

Question 1: Pollution Pile-ups (25 points, 25 minutes)

Dr. Green believes that pollution levels should go down the further you are from a major traffic source such as a freeway but she is not sure how rapidly they drop off. She has therefore conducted a small study in which she has recorded Y, the total particle concentration (in units of thousands of particles per cubic centimeter) at a variety of distances, X, (measured in meters) from the 47 freeway, the major road into Los Seraphim. A STATA printout of the simple linear regression for her data, and some corresponding confidence and prediction intervals for Y are shown below. Use them to answer the questions that follow.

```
. reg pollution distance

      Source |       SS       df       MS              Number of obs =      26
-------------+------------------------------           F(  1,    24) =   21.47
       Model |  23400.0003     1  23400.0003           Prob > F      =  0.0001
    Residual |  26160.1198    24  1090.00499           R-squared     =  0.4722
-------------+------------------------------           Adj R-squared =  0.4502
       Total |  49560.1201    25   1982.4048           Root MSE      =  33.015

------------------------------------------------------------------------------
   pollution |      Coef.   Std. Err.      t     P>|t|    [95% Conf. Interval]
-------------+----------------------------------------------------------------
    distance |        -.2   .0431655    -4.63   0.0001   -.2890891   -.1109109
       _cons |        150   12.58478    11.92   0.0000    124.0263    175.9737
------------------------------------------------------------------------------

. adjust distance = 725, se ci
------------------------------------------------------------
      All |         xb        stdp          lb          ub
----------+-------------------------------------------------
          |          5   (21.5016)   [-39.3772    49.3772]
------------------------------------------------------------
     Key:  xb         =  Linear Prediction
           stdp       =  Standard Error
           [lb , ub]  =  [95% Confidence Interval]

. adjust distance = 725, stdf ci
------------------------------------------------------------
      All |         xb        stdf          lb          ub
----------+-------------------------------------------------
          |          5   (39.3996)   [-76.3167    86.3167]
------------------------------------------------------------
     Key:  xb         =  Linear Prediction
           stdf       =  Standard Error (forecast)
           [lb , ub]  =  [95% Prediction Interval]
```

Part a (7 points)

Dr. Green believes there is a negative relationship between pollutant level and distance from the freeway. Perform the hypothesis test she would use to prove her theory. Give the null and alternative hypotheses mathematically and in words with a justification of your choice, obtain the p-value and give your real-world conclusions using $\alpha = .05$.

Solution: A negative relationship between distance and pollution means that as you get further away from the freeqay, the pollution level (i.e. particle concentration) gets lower. This corresponds to the slope in our simple linear regression model being negative and is a 1-sided test. Since Dr. Green wants to prove that there is a negative relationship, that is her alternative hypothesis:

$H_0 : \beta_1 \geq 0$–there is a positive or no relationship between distance from the freeqay and particle concentration.
$H_A : \beta_1 < 0$–there is a negative relationship between distance from the freeway and particle concentration.

The STATA printout gives the p-value for a two-sided test of $H_0 : \beta_1 = 0$ versus $H_A : \beta_1 \neq 0$. To get the p-value for our 1-sided test, we must divide this p-value in half, getting $.0001/2 = .00005$. This p-value is definitely less than our significance level of $\alpha = .05$ so we reject the null hypothesis and conclude that there is a significant negative relationship between distance from the freeway and particle concentration. As we get further away from the freeway, air quality improves (hardly a surprise!)

There were a couple of mistakes that showed up fairly frequently in this problem. One was that people stated the hypotheses in terms of the sample slope, b_1, rather than in terms of the population value, β_1. For a particular sample, b_1 is a fixed known number–in this case $b_1 = -.2$. You can't have an hypothesis about it. Rather you are using the sample value to figure what the possible values of the population slope are. That is what your hypotheses are about. Second, a number of people tried to state the hypotheses in terms of means, μ. This might make sense if we had divided our sample into two groups, close to the freeway and far from the freeway, and had used an indicator variable. However it doesn't make sense for a continuous predictor. We are not interested here in the mean value at a particular distance from the freeway but rather in how the average pollution level changes continuously with distance. Finally, many people forgot to divide the p-value by 2 for the 1-sided test and forgot to give a justification for their chocie of hypotheses.

Part b (4 points)

Find the correlation between pollution level and distance from the freeway. Does there appear to be a strong relationship between these variables? Explain briefly.

Solution: In a simple linear regression, the correlation is the square root of R^2, the percentage of variability explained, **with the appropriate sign attached!** Here $R^2 = .4722$ and we have a negative relationship (per part (a)) so the correlation is $r = -\sqrt{.4722} = -.687$. Many people forgot the negative sign. Correlation has a direction while R^2 does not. Whether this is a strong

correlation is a bit of a judgment call. According to the rough (very rough!) rules of thumb I gave in class, a correlation that is greater than .7 in absolute value is strong, so we are between a medium and a strong correlation by those standards. However in this context, I think -.687 is pretty good. There are many other factors besides distance from the freeqay that affect pollution including wind, temperature, proximity to other pollutant sources, geography (that may tend to trap or dissipate pollution) and so on.

Part c (3 points)

Public Health guidelines suggest that exposure to fine particle pollution levels of more than 50,000 particles per cubic centimeter is unhealthy. According to this model, what is your best estimate of how far you have to get from the freeway before the fine particle concentration drops to this level?

Solution: A pollution level of 50,000 particles/cm^3 corresponds to a value of $Y = 50$ since Y is measured in thousands of particles per cubmic centimeter. We want to find the X value that gives us this value of Y. Thus we have

$$\hat{Y} = 50 = b_0 + b_1 X = 150 - .2X$$

or

$$X = \frac{50 - 150}{-.2} = 500$$

Thus if we are at a distance of 500 meters from the freeqay, the expected particle concentration is 50,000/cm^3. Since the relationship between pollution and distance is negative, if we are even further away from the freeqay, the pollution level will also be below the 50,000 level.

The most common mistakes on this part of the problem were to forget to put Y in units of thousands of particles (i.e. plugging in 50,000 instead of 50) or to treat 50 as if it were the X variable rather than the Y variable.

Part d (6 points)

The Los Seraphim Unified School District is about to build a large number of new elementary schools. It has decided that it will build them no closer than $X_0 = 725$ meters from the 47 freeway. The district Superintendent wants to be sure that the new schools schools will be under the critical exposure level from part (c). Can you be sure that 95% of schools at this distance will be under the threshold? Explain which interval from the printout you are using to address this question and why. (No calculations are required.)

Solution: This was one of the most missed parts of the exam because it makes a subtle distinction about the difference between confidence intervals and prediction intervals for Y. Recall that a CI gives you a range of values which you are fairly sure (here 95%) will include the **average** value of Y at a given X while a prediction interval gives you a range of values which you are fairly sure will include any **individual** value of Y. To say you are 95% sure that a PI will include the value

of Y for a particular or individual subject at a given X means that 95% of the time you would expect your Y value for an individual to fall in the interval–i.e. 95% of individual Y values at a particular X should land in the range specified by the prediction interval. We have no idea what fraction of individual Y values will lie in the CI for the average Y at a given X. We just know that the average value is highly likely to be in the interval. Now let's look at the problem statement. The superintendent wants the kids at the schools to be safe from harmful pollution. It doesn't say he wants the **average** school to be safe. That would mean there were lots of schools that wouldn't be safe! What it says is that he wants 95% of schools–that is the vast majority of the **individual schools** to be safe. This means we need a **prediction interval**.

The second part of the problem asked whether we could be sure that the schools would be below the threshold of 50,000 particles/cm^3 or Y = 50 if we built the schools a distance of at least 725 meters from the freeway. According to the STATA printout, the PI for Y at $X = 725$ is [-76.32, 86.32]. Naturally we can't have a negative pollution level so what this interval really tells us is that we are 95% sure each individual school will have a pollution level below 86,320. Since the interval includes values well above the safe threshold of 50,000, the superintendent can NOT be sure that most of the individual schools will be safe.

A couple of notes: First, a number of people tried to say that you couldn't be sure that the schools were safe because the PI (or CI) included the value 0. While 0 is often an important value for us (e.g. $\beta = 0$ in a regression means no relationship), it is not an important value in this problem. We are talking about a PI for Y, the actual pollution level. If $Y = 0$ that means there is no pollution so the school is certainly safe in that respect. The critical value we are interested in is $Y > 50$. Some people calculated the expected pollution at $X = 725$ meters, which is 5,000 particles/cm^3, and tried to claim that the schools were safe because this value is well below 50,000. This is the best estimate for the schools and certainly looks good. However, individual schools will vary about that predicted value and we are concerned with all (or most) of the schools being safe, not just with our best guess about whether one will be safe. People also tried to say that since the value $Y = 5$ was in the PI this meant that our results were OK. This doesn't make sense either. The PI (and CI) for Y both take the form $\hat{Y} \pm .....$ This value will **always** be in the interval, right in the center. It's inclusion in the interval doesn't tell you anything. Some people also tried to say that because we were interested in a particular or individual distance, $X = 725$, we needed a PI. You are **always** interested in a particular value of X–you can't make a prediction at more than one value at a time. The key distinction is whether you are interested in an individual vs average **Yvalue** at that particular X value. Finally, a lot of people argued that because we were looking at "many" schools we had to use a CI. There is a difference between caring about the individual pollution levels at many schools and caring about the average pollution level at many schools. Only the second one uses a CI. The PI gives the range in which each individual school at a given distance from the freeway is likely to be–but it is the same range for each school so we can get information about many schools from it.

Part e (5 points)

Cautious Connie suggests it would be even safer to build the new schools a distance of 1000 meters from the freeway. Find the estimated particle concentration at this distance. Does your answer make real-world sense? Discuss briefly why what has happened makes sense in the context of the

problem.

Solution: To get the estimated particle concentration we just plug into the regression equation:

$$\hat{Y} = b_0 + b_1 X = 150 - .2(1000) = 150 - 200 = -50$$

Thus at a distance of 1000 meters from the freeway, our model predicts a particle concentration of -50,000.cm^3. This obviously makes no sense. You can't have a negative particle concentration. What has gone wrong is that we are predicting outside the range where this model is valid. (I actually originally had summarize statements for both the X and Y variables showing the the range of values used in the data set but I took them out because they weren't needed and added a lot of extra text to read–I didn't use any values greater than X = 500.) That far almost everybody got, but there is one more key point, namely why does the model become invalid outside the range I used to create it? The reason extrapolation is dangerous is that the nature of the relationship (shape of the plot) can **change** for different values of X and if you don't have data in a given range you won't see the change. (If the nature of the relationship stays the same across all X then the extrapolation will not be as bad though it will still run into problems because the true and sample lines will grow further apart as you get further from the center of your data.) In this case the relationship between distance and pollution, practically speaking, has to **level off** if you get far enough away from the freeway, because the minimum possible value is 0. It is not possible for a negative linear relationship to hold indefinitely although it may work quite well in the limited range of data used to build the model. We'll see later how to fit a more appropriate curvilinear model to this data which will account for that leveling off.

Question 2: In and Out Statistics (That's What This Math Test's All About...20 points, 20 minutes)

Next Dr. Green has decided to compare particle concentrations inside the students' classrooms to outside on the playground. She has taken $n_I = 11$ indoor measurements and $n_O = 11$ outdoor measurements and found $\bar{Y}_I = 25$, $s_I = 10$, $\bar{Y}_O = 30$, $s_O = 10$. (Note: As before the pollution levels are in thousands of particles per cm^3. You may assume for purposes of this problem that the various measurements are independent).

Part a (2 points)

What is the pooled estimate of the standard deviation for the particle level concentration? Show your work or explain your reasoning.

Solution: There's a quick way and a slow way to answer this. The pooled estimate of the standard deviation is based on combining the variability of points within each of the groups. Here the standard deviation within both groups is the same, $s = 10$, so the pooled estimate must also be $s_p = 10$. The long wat to do this is to use the formula

$$s_p = \sqrt{\frac{(n_I - 1)s_I^2 + (n_O - 1)s_O^2}{n_I + n_O - 2}} = \sqrt{\frac{(11-1)10^2 + (11-1)10^2}{11 + 11 - 2}} = \sqrt{\frac{2000}{20}} = \sqrt{100} = 10$$

Part b (7 points)

Find a 95% confidence interval for the difference between pollution level inside and outside the classrooms. Based on this interval can you be sure there is a difference? If not, why not? If so, which location seems more polluted?

Solution: The confidence interval for a two-sample difference in means is

$$\bar{Y}_O - \bar{Y}_I \pm t_{\alpha/2, n_I + n_O - 2} s_p \sqrt{\frac{1}{n_I} + \frac{1}{n_O}}$$

We are given $\bar{Y}_O = 30, \bar{Y}_I = 25, n_O = n_I = 11$. Since we want a 95% confidence interval we therefore need $t_{.025,20} = 2.086$. Combining these numbers with our answer to part (a) we have

$$30 - 25 \pm (2.086)(10)\sqrt{\frac{1}{11} + \frac{1}{11}} = [-3.89, 13.89]$$

Based on this interval we can **not** be sure there is a difference in pollution levels inside and outside the classroom. The CI for the difference in means includes 0, so it is possible that there is no difference. Note that this does not mean we have **proved** there is no difference, just that we do not have sufficient evidence to show there is a difference with 95% confidence. In fact, our best estimate is that the pollution level is 5000 particles/cm^3 higher outdoors than indoors and a much larger portion of the interval is above 0 than below it so my best guess is that pollution is worse outside but it is possible, given our sample size and observed values, that that result is due to chance.

Part c (4 points)

Give an estimate for the effect size for the difference between inside and outside measurements. Is this a large effect size? Explain.

Solution: The effect size for a two-sample comparison of means (assuming the groups have equal variance) is given by the difference in means divided by the pooled estimate of the standard deviation:

$$d = \frac{\bar{Y}_O - \bar{Y}_I}{s_p} = \frac{30 - 25}{10} = .5$$

The means differ by half a standard deviation. The second part of the question asked whether this was a large effect size. According to the conventions of Jacob Cohen (followed nearly as slavishly in the social sciences as Fisher's recommendation to use $\alpha = .05$ as the significance level for CIs and hypothesis tests!) this is a standard medium sized effect. However, just as with our rules of thumb for what constitutes a large correlation, what counts as a large effect really depends on context and field. Here part of what we would want to know whether a difference of 5000 particles/cm^3 between indoor and outdoor pollution levels was meaningful for the school childrens' health. The "standardized" effect size of Cohen tells us that it amounts to about half a standard deviation in terms of the normal variability of pollution but is possible for that to be clinically important or unimportant. As long as you gave a reasonable discussion of the issues you should have gotten at least partial credit for this whether you said medium or not.

Part d (7 points)

Before carrying out her study, Dr. Green carried out the power calculation shown below based on the assumption that the true mean difference between inside and outside was $\mu_0 - \mu_I = 10$ and the standard deviation in both locations was $\sigma = 10$.

```
. sampsi 20 30, sd(10) n(11)

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 =       20
            m2 =       30
           sd1 =       10
           sd2 =       10
sample size n1 =       11
            n2 =       11
         n2/n1 =     1.00
```

```
Estimated power:

         power =   0.6500
```

(i) Do you believe the study was adequately powered? Explain.
(ii) Name two things Dr. Green could practically have done to increase her power and briefly discuss their pros and cons.

Solution: This study was **not** adequately powered. Usually the target is 80% power for a 2-sided test with $\alpha = .05$ (an arbitrary convention, as we have discussed, but very standard.) Here we have powered for a two-sided test with $\alpha = .05$ but the sample size chosen is only large enough to give us 65% power. This means that even if our assumptions about the true mean difference and standard deviations are correct, 1 time out of 3 our study will fail to detect a group difference.

There are a number of things that increase power. One is to increase the sample size. The more information you have, the better your chance of correctly identifying whether or not there is an effect. This is a good idea (the statistician will ALWAYS tell you to get a bigger sample!) if she can. However it may be expensive or time-consuming to take extra measurements. (In some situations, especially with a rare disease, it may not be possible to recruit additional subjects in the time frame of the study. However that is probably not the case here.) Another way to increase power is to raise the significance level, α. The easier it is to reject the null hypothesis overall, the easier it will be to reject when the null hypothesis is false which is what power is. This has the advantage of being really easy! Dr. Green just has to change one number in her proposal. No extra measurements are involved. However there is a significant downside, namely that increasing α increases the chances of a Type I error, in this case the probability of concluding there is a difference between inside and outside pollution levels when there isn't really. Another option is to perform a one-sided test rather than a one-sided test. If you had strong evidence from the literature or a theory that provided a good argument for the pollution being higher in one location or the other, then you might be able to get away with this. Some times only one direction will actually be of interest. However, funding agencies and reviewers usually insist on a two-sided test anyway. The reason most commonly given is that you can't be sure which way the difference will go and you don't want to miss a significant result in the opposite direction. An actually better reason is the following. All our calculations for hypothesis tests, including power, are based on an underlying assumption of normality. However if the true distributions are skewed the normal approximation can seriously misrepresent the probabiltiy of being in the tail of the distribution. If you do a one-tailed test that can have a big effect on your p-value. However if you do a two-tailed test, the errors from the two tails tend to cancel each other out and your estimate of the p-value is more accurate.

The three approaches above are probably Dr. Green's best bet for increasing her power. There are other things that result in increased power, like smaller standard deviations, or bigger differences between the groups, but those things are not under Dr. Green's control so practically she can't hope to change them to increase her power. She could make different **assumptions** about what the means and SDs were but if her assumptions do not seem plausible that will also get her in trouble.

Part e (Optional Bonus–Up to 5 points)

A t-test can be thought of as a simple linear regression where the only predictor is a single indicator variable. Let Y be particle concentration and let $X = 1$ if the measurement is taken outside and $X = 0$ if it is taken inside. Our model is $Y = \beta_0 + \beta_1 X + \epsilon$.

(i) Give brief real-world interpretations of β_0 and β_1 and what you think the best estimates of these quantities, b_0 and b_1, should be based on the given data.

(ii) Fill in the ANOVA table corresponding to this simple linear regression showing your work or explaining your reasoning.

Analysis of Variance

Source	SS	df	MS	F	Prob > F
Regression	137.5	1	137.5	1.375	> .2_
Error	2000.0	20	100.0		
Total	2137.5	21			

Solution: In a simple linear regression, the intercept is the average value of Y when $X = 0$. Here X is an indicator and when it equals 0 that means the measurement was taken **inside**. Therefore β_0 is the average pollution level inside the classrooms. For an indicator variable, the corresponding coefficient gives the average difference in Y between observations that do and do not have the characteristic measured by the indicator. Here $X = 1$ if the measurement is taken outside so β_1 is the average difference in pollution level between measurements taken outside and measurements taken inside. We know that the best estimate of the average indoor and outdoor measurements are–they are given by the group means at the beginning of the problem. Thus $b_0 = \bar{Y}_I = 25$ is the best estimate of the inside pollution levels and $b_1 = \bar{Y}_O - \bar{Y}_I = 30 - 25 = 5$ is the best estimate of the average difference between inside and outside pollution levels. Note that $b_0 + b_1(1) = 30 = \bar{Y}_O$ is the best estimate of the average pollution levels outside.

The filled in ANOVA table is given above. The degrees of freedom for a simple linear regression are always 1, $n - 2$ and $n - 1$. Since we have $n_I = n_O = 22$ we must have 1, 20 and 21 degrees of freedom respectively. There are several ways we can fill in the rest of the table. First mean squared error must be 100 because MSE is the average squared error of points about the values predicted by the model. The predictions given by the models are just the location means, so MSE is the average squared error of points about the means for each location. But we were told that the standard deviation for each lcoation was 10 and this the variance or squared error at each location must have been 100. Alternatively you can think of this as an ANOVA with two groups in which case MSE = MSW which is just the square of the pooled estimate of the standard deviation which we saw was 10 in part (a). It follows that SSE = 2000 since SSE = MSE*df = 100*20. For the next part one approach is to recall that in a simple linear regression the F statistic is the square of the t-statistic for β_1. But here the test for β_1 amounts to a test for the difference in group means. This is just the two-sample t-test that parallels the confidence interval from part (b). We would have

$$F = t_{obs}^2 = (\frac{30-25}{10\sqrt{1/11+1/11}})^2 = 11/8 = 1.375$$

Since F = MSR/MSE, MSR = F*MSE = 100*1.375 = 137.5. In a simple linear regression SSR = MSR so SSR = 137.5. Finally, SST = SSR + SSE = 137.5 + 2000 = 2137.5. This completes the ANOVA table.

Alternatively, one could use the ANOVA formulas for SSB and SSW recognizing this as a two-group ANOVA. The grand mean is $\bar{Y} = (25+30)/2 = 27.5$, the group means are $\bar{Y}_I = 25$ and $\bar{Y}_O = 30$ and the two sample variances are $s_I^2 = s_O^2 = 100$, and the group sizes are $n_I = n_O = 11$. This we have

$$SSB = \sum_j n_j(\bar{Y}_j - \bar{Y})^2 = 11(25-27.5)^2 + 11(30-27.5)^2 = 137.5$$

Similarly, for the within group variation we have

$$SSW = \sum_j (n_j - 1)s_j^2 = (11-1)(100) + (11-1)100 = 2000$$

Once these values are filled in the rest of the table is is easy.

Finally, we can not get the exact p-value without STATA. However by taking the square root of the F statistic we get the t-statistic for testing $\beta_1 = 0$ vs $\beta_1 \neq 0$. The value is $t_{obs} = 1.17$. This should register for you as a very small value of t. If we check our t-table with 20 degrees of freedom, we find that $t_{20,.10} = 1.325$. This is the one-tailed critical value. Since $t_{obs} = 1.17 < 1.325$ it follows that the 1-sided p-value is greater than .1 and the 2-sided p-value is greater than .2.

Question 3: Location, Location, Location (30 points, 35 minutes)

Dr. Green has decided to compare the pollution levels in several different locations and conditions. Specifically, she decides to take measurements in the parking (P) lot at morning drop-off, in the play yard at afternoon recess (R), in the cafeteria (C) at lunch time, and in two classrooms right before lunch, one with the windows closed and air conditioner running (A) and one with the windows open (W). An ANOVA printout for her data is shown below. Use it to answer the following questions. As before assume Y is in thousands of particles per cm^3 and measurements are independent.

```
. oneway particle location, tabulate bonferroni

                |         Summary of particle
       location |        Mean   Std. Dev.     Freq.
 ---------------+------------------------------------
 A (class, AC)  |        14.2      7.1         25
 C (cafeteria)  |        19.1      6.9         25
 P (parking lot)|        37.9      6.7         25
 W (window open)|        36.8      5.1         25
 R (play yard)  |        41.5      9.0         25
----------------+------------------------------------
          Total |        29.9     13.1        125

                        Analysis of Variance
    Source              SS      df       MS         F   Prob > F
------------------------------------------------------------------------
Between groups      15225.6      4   3806.4     76.13     0.0000
 Within groups       6000.0    120     50.0
------------------------------------------------------------------------
    Total           21225.6    124    171.2

           Comparison of particle by location   (Bonferroni)
Row Mean-Col Mean|
        p-value  |          A          C          P          W
        ---------+-------------------------------------------
              C |        4.9
                |      0.161
                |
              P |       23.7       18.8
                |      0.000      0.000
                |
              W |       22.6       17.7       -1.1
                |      0.000      0.000      1.000
                |
              R |       27.3       22.4        3.6        4.7
                |      0.000      0.000      0.723      0.192
```

Part a (5 points)

Is there overall evidence of a difference in pollution levels at the various locations? Justify your answer with an appropriate test, giving the mathematical hypotheses and your real-world conclusions using $\alpha = .05$.

Solution: Here we are being asked to perform an overall F test. Our hypotheses, using the notation of classical ANOVA, are

$H_0 : \mu_A = \mu_C = \mu_P = \mu_W = \mu_R$–the mean pollution level is the same in all 5 of the locations (parking lot, yard, cafeteria, window open classroom, window closed classroom.)
H_A : Not all the μ_j's are equal. The pollution level is related to which location you are at.

From the printout, the test statistic is a whopping $F = 76.13$ and the corresponding p-value is 0 to as many decimal places as STATA gives. We therefore reject the null hypothesis at $\alpha = .05$ (or any other significance level down to .0001) and conclude that pollution level does vary by location on the school grounds. Note that you only needed the mathematical hypotheses and a real-world conclusion based on the p-value. I included the other details for completeness. By "real-world" I mean that I want the conclusion in the context of the problem (pollution and location) rather than just saying "we reject."

Part b (5 points)

According to the printout, after using the Bonferroni adjustment for multiple comparisons, which pairs of locations do **not** have significantly different particle concentrations? Use this information to describe as precisely as you can the ordering of the locations in terms of air quality for the students.

Solution: To get it's Bonferroni "adjusted" p-values, STATA takes the raw p-values and multiplies them by the number of tests so that they can be directly compared to the overall significance level, α. Thus we just need to compare the p-values on the printout to $\alpha = .05$. All the comparisons have p-values smaller than this except the air-conditioned classroom (A) vs the cafeteria (C), the parking lot (P) versus the open-windowed classroom (W), the parking lot vs the play yard (R) and the window-open classroom vs the play yard. Thus it seems that we really have two groups of pollution levels–the air-conditioned classroom and the cafeteria are better (lower pollution) that the parking lot, yard and window open classroom. Our best guess is that the order (from lowest to highest pollution) is A, C, W, P, R but we can only be 95% sure that the (A,C) pair is significantly better than the (W,P,R) triplet.

Part c (4 points)

Which pairs of locations would have been significantly different **without adjusting for multiple testing.** Briefly explain your reasoning.

Solution: As noted in part (b), STATA obtains it's adjusted p-values by taking the original individual p-values and multiplying them by the number of tests. Since there are 10 tests here (5 choose 2) we can get back the original p-values by dividing the given Bonferroni p-values by 10. Obviously the 6 differences that were significant after adjustment will remain significant. The A

vs C comparison (air-condition class vs cafeteria) must have had an original p-value of .161/10 = .0161 so it would have been significant without adjustment for multiple comparisons. Similarly, the W vs R (window open class vs play yard) comparison would have been significant with an original p-value of .0192. However the P vs R (parking lot vs play yard) p-value would have been .0723 which is not significant. We actually can't be sure what the P vs W (parking lot vs window open classroom) p-value would have been because STATA truncates it's adjusted p-values at 1 and P vs W has reached that level. What we do know is that the adjusted p-value was at least 1 so the unadjusted p-value would have been at least .1 which is not significant. Thus 8 out of the 10 comparisons were significant before adjusting for multiple comparisons.

Part d (12 points)

Dr. Green is interested once again in comparing inside air quality to outside air quality.

(i) Write down the linear combination in which she is interested.
(ii) Give your best estimate for the linear combination and its standard error based on this data.
(iii) Perform the hypothesis test Dr. Green would use to show that the average air quality inside the buildings is at least 10,000 particles/cm^3 better than outside.

Solution: My intention when designing the problem was that the parking lot and play yard counted as outdoor locations and that the two classrooms and cafeteria counted as indoors. However some people interpreted the classroom with the open window as outside since it was getting outdoor air. In fact this classroom is probably halfway in between so we accepted either version. I present the calculations as if the window-open classroom is an indoor location.
(i) We wish to compare the mean of the three indoor locations to the mean of the two outdoor locations. The linear combination corresponding to this comparison is

$$LC = \frac{\mu_W + \mu_C + \mu_A}{3} - \frac{\mu_P + \mu_R}{2}$$

Note that the linear combination is just the expression for the difference between the groups. It does **not** include the "=0" piece. That is a particular value of the difference that we might be interested in checking but it is not formally part of the linear combination itself.

(ii) To estimate the linear combination we substitute the sample group means for the population group means in the expression above:

$$\hat{LC} = \frac{\bar{Y}_W + \bar{Y}_C + \bar{Y}_A}{3} - \frac{\bar{Y}_P + \bar{Y}_R}{2} = \frac{36.8 + 19.1 + 14.2}{3} - \frac{37.9 + 41.5}{2} = -16.333$$

Note that you could also reverse the order (taking the outdoor average minus the indoor average) which would result in $\hat{LC} = 16.33$. All it does is change the sign. (If you grouped the window-open classroom with the play yard and the parking lot the indoor measurements were 22.083 lower than the outdoor measurements.)

The estimated standard error of a linear combination is given by

$$s.e.(\hat{LC}) = \sqrt{MSW * \sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

From the ANOVA table we have MSW = 50, there are $n_j = 25$ measurements at each location, and the constants in the linear combination are $1/3, 1/3, 1/3, -1/2, -1/2$ as I have written it. The correspnding standard error is

$$s.e. = \sqrt{50(\frac{1}{25})(3 * (\frac{1}{3})^2 + 2 * (\frac{-1}{2})^2)} = \sqrt{2 * (\frac{1}{3} + \frac{1}{2})} = 1.29$$

Note that the standard error is the same no matter how you group the locations since either way you have 3 of one type and 2 of the other.

(iii) Dr. Green wants to test the hypothesis that air quality inside the classrooms is 10,000cm^3 better, i.e. **lower** inside the buildings than outside. Since my linear combination was set up to subtract the outdoor locations from the indoor locations and particle concentrations are in thousands, this means I want to prove that my linear combination is **less** than -10, so that is my alternative hypothesis :

$H_0 : LC \geq -10$–the pollution levels at the inside locations are not at least 10,000 particles/cm^3 better than in the outdoor locations.
$H_A : LC < -10$–the pollution levels for the indoor locations are at least 10,000 particles/cm^3 lower than for the outdoor locations.

The test statistic is

$$t_{obs} = \frac{\hat{LC} - LC_{H_0}}{s.e.(\hat{LC})} = \frac{-16.33 - (-10)}{1.29} = -4.91$$

Since we are doing a 1-sided test with an alternative of "less than," and we have $n - k = 120$ degrees of freedom associated with our within group variability, our p-value is

$$p - value = P(t_{120} \leq -4.91) = .00000145$$

I got the exact p-value from STATA which you couldn't do during the exam. The best you can do from the t-table is that $t_{120,.005} = 2.617$. Since our test statistic is bigger than this in absolute value, the p-value must be less than .005. In any case, we clearly reject the null hypothesis at $\alpha = .05$ and conclude that the pollution level is at least 10,000 particles/cm^3 better inside than outside. (You get this conclusion even more strongly if you group the open window classroom with the outdoor measurements since its mean was more like the other outdoor locations than the other indoor locations.)

Part e (4 points)

Suppose that Dr. Green had wanted to include the test of the linear combination from part (d) in her correction for multiple comparisons. Name two approaches she could have used and say very briefly (no more than a sentence) why each is appropriate.

Solution: There are many possible answers to this question. One approach is to continue using the Bonferroni correction since it applies to any set of tests. You will simply have to divide α by 11 rather than 10 to account for the extra test. Another approach is to use the Scheffe' method which specifically allows you to simultaneously test all possible linear combinations. Since pairwise mean comparisons are linear combinations as is the test from part (d), Scheffe applies. You could also use the Holm, False Discovery Rate or the Omnibus approaches which are all valid for any collection of tests. What you can not do is use a mthod like Tukey, SNK or the Fisher LSD approach because these apply only to pairwise comparisons of means and the test in part (d) is not a pairwise comparison of means. We accepted any reasonable pair of tests.

Question 4: A Fine Model For Fine Particles (25 points, 30 minutes)

Dr. Green realized that many factors besides location might affect the air quality at the schools she was studying. Therefore in the data set collected for Question 3 she also recorded X_1, the distance of the measurement from the freeway or other major traffic source (in meters), X_2, the temperature at the time the measurement was taken (in degrees Fahrenheit, and an indicator for whether or not it was raining at the time of the measurement: $X_3 = 1$ for rain and $X_3 = 0$ for no rain. She has also used a set of 4 indicators for the different locations: $X_4 = 1$ if the measurement is in the parking lot in the morning, $X_5 = 1$ if the measurement is in the cafeteria at lunch, $X_6 = 1$ if the measurement is in a classroom before lunch with the windows closed and $X_7 = 1$ if the measurement is in a classroom with the windows open before lunch. The play yard at afternoon recess is the reference group with $X_4 = X_5 = X_6 = X_7 = 0$. Dr. Green's multiple regression printout is shown below. Use it to answer the following questions.

```
. regress particle distance temperature rain parking cafeteria closed open

      Source |       SS       df       MS              Number of obs =     125
-------------+------------------------------           F(  7,   117) =   52.43
       Model |  18350.6        7    2621.5             Prob > F      =  0.0000
    Residual |   2875.0      117      25.0             R-squared     =  0.8646
-------------+------------------------------           Adj R-squared =  0.8564
       Total |  21225.6      124     171.2             Root MSE      =   5.000

------------------------------------------------------------------------------
     particle |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
     distance |     -0.05    0.0125       4.00   .0001     [-0.075,     -0.025]
  temperature |      0.40    0.1500       2.67   .0086     [ 0.103,      0.697]
         rain |    -15.00    5.0000      -3.00   .0033     [-24.90,     -5.098]
      parking |     10.00    2.0000       5.00   .0000     [ 6.039,     13.961]
    cafeteria |     -5.00    2.0000       2.50   .0138     [-9.951,     -0.049]
window closed |    -12.00    2.0000      -6.00   .0000     [-15.96,     -8.039]
  window open |      0.00    1.5000       0.00  1.0000     [-2.971,      2.971]
        _cons |     60.00    5.0000      12.00   .0000     [50.098,     69.902]
------------------------------------------------------------------------------
```

Part a (7 points)

Give units and real-world interpretations for b_2 and b_3, the coefficients of the temperature and rain variables. Make sure your answers incorporate the numerical values of the coefficients.

Solution: The coefficient of the temperature variable is $b_2 = .4$. For a continuous predictor, the coefficient gives the change in Y associated with a one unit change in X, **assuming all the other variables are held fixed.** Here X_2 is in degrees Fahrenheit and Y is in thousands of particles per cm^3 so this says that all else equal, on average a 1 degree increase in temperature is associated with a .4*1000 = 400 particle/cm^3 increase in poluution levels. Another way of saying this is that if I take pollution measurements at the same location (parking lot, yard, etc.) at the same distance from a freeway under the same rain conditions but one measurement is taken when the temperature is a degree hotter than when the other measurement is taken then the pollution level for the hotter measurement will on average be 400 particles/cm3 higher.

The coefficient for the rain variable is $b_3 = -15$. For an indicator variable the coefficient gives the difference in Y between observations that do and do not have the characteristic, assuming all other variables are held fixed. Here this means that, all else equal, on average pollution levels are 15,000 particles/cm^3 lower when it rains than when it doesn't rain. Note that it is very important to indicate what groups you are comparing. Many people said "lower when raining" without saying "lower compared to what." In this case it is pretty obvious since there are only two categories and we weren't too mean about it. However it is easy to get confused when there are more categories. Many people got this tangled up in part (d) of this problem for instance so do be explicit.

The most common mistake on this part, other than not explicitly noting the reference group for an indicator interpretation was that people forgot to say "all else equal" or "holding all the other variables fixed." This is critical in multiple regression. You must do all interpretations accounting in some way for the presence of the other variables. If you do not hold the other variables fixed then you can not tell what part of the resulting effect is due to the variable of interest.

Part b (5 points)

Use the model to predict the pollution level in the parking lot on a rainy day when it is 60 degrees at a school that is 1000 meters from the nearest major freeway.

Solution: To make a prediction we just plug in the various X values, being careful of our units and the various indicator variables. We are told the distance is 1000 meters so $X_1 = 1000$, the temperature is 60 degrees so $X_2 = 60$, it is raining, so $X_3 = 1$, and we are taking the measurement in the parking lot so $X_4 = 1$ and $X_5 = X_6 = X_7 = X_8 = 0$. Don't forget to include the itnercept in your prediction! We have

$$\hat{Y} = 60 - .05(1000) + .4(60) - 15(1) + 10(1) - 5(0) - 12(0) + 0(0) = 29$$

Since Y is in thousands we have that the predicted pollution level in the parking lot on a rainy day with a temperature of 60 degrees Fahrenheit, 1000 meters from the freeway, is 29,000 particles/cm^3.

Part c (5 points)

The Los Seraphim Times reports that the temperature today is going to be 10 degrees hotter than yesterday but otherwise conditions around the city will be the same. Mrs. Jones wants to know how this is going to affect the air quality at her daughter's school. Give a 95% confidence interval for the average difference in air quality she should expect between today and yesterday, briefly explaining your reasoning.

Solution: What Mrs. Jones is interested in is the change in pollution level associated with a 10 degree change in temperature, assuming everything else is held fixed. We know that everything else is being kept constant because we are told that the weather conditions other than temperature will be the same as yesterday (i.e. the rain status won't change), Mrs. Jones' daughter will be going to the same school so the distance from the freeway will be the same, and we can compare pollution levels at any location at the school to those at the same location on the previous day. Thus we are talking about the effect of a change in X_2 and our answer will be based on the corresponding coefficient, b_2. Now b_2 gives the change in Y (pollution level) associated with a 1 degree change in temperature, all else held fixed. The confidence interval for b_2 is [.103, .697] which means that a 1 degree change in temperature is associated with an increase in pollution levels of between 103-697 particles per cm^3 with our best estimate being a change of 400 particles. To get the change in pollution associated with a 10 degree change in temperature we simply multiple the estimate and interval by 10, giving us an increase in pollution level of between 1030-6970 particles/cm^3 with a best estimate of 4000 particles/cm^3. Taking the existing estimate/CI and multiplying them by 10 is the simplest way to answer this question. If we go up by β_2 for a 1 degree change in temperature we go up $10\beta_2$ for a 10 degree change if a linear model is correct. This applies to both our average, best case and worst case estimates of β_2.

Many people looked at the best estimate of 400 particles/cm^3 per degree, converted this to 4000 particles/cm^3 for 10 degrees and then tried to calculate the corresponding CI by hand. This works, but you have to remember that in this case you must multiply the standard error by 10 as well as the estimate. Most people who took this approach just used the standard error for b_2 instead of the standard error of $10b_2$. For any constant, a, and random variable V, $SD(aV) = aSD(V)$. The other common mistake people made was to try to turn this into a confidence interval or prediction itnerval for Y. This would make sense if we wanted an interval for the actual pollution level, but Mrs. Jones wants an interval for the **change** in pollution level from yesterday, not the pollution level itself. In addition, I never taught you how to do the hand calculation of a CI or PI for Y for a multiple regression so this is not a calculation you would have been able to do–always a good hint that it is not what I am looking for :)

Part d (4 points)

Based on this model, does there appear to be a significant difference (using $\alpha = .05$) between pollution levels in the parking lot and in the play yard? Briefly justify your answer.

Solution: Yes, there does appear to be a significant difference in pollution level between the play yard and the parking lot. In this model, the play yard was the reference location and we had indicators for all the other locations. Thus β_4, the coefficient for the indicator of the parking lot

location is precisely the difference in pollution level between the parking lot and the play yard. From the printout we see that β_4 is significantly different from 0, looking either at the p-value of .0000 or the confidence interval of [6.039, 13.961]. In fact, since this interval is entirely above 0 we can be sure that the pollution level in the parking lot is **higher** than the pollution level in the play yard.

Part e (4 points)

Your answer to part (d) is different to what was found in Question 3. Explain briefly what you believe has happened.

Solution: In Question 3, we failed to find a significant difference between pollution levels in the parking lot and play yard and in fact our best estimate was that the pollution levels were on average higher in the play yard (at 41. 5) than they were in the parking lot (at 37.9). What has changed from Question 3 is that we are fitting a multiple linear regression instead of an ANOVA and in particular are taking into account the effects of distance from the freeway, temperature, and whwther or not it is raining. That our answers changed from when we were considering location only implies that accounting for these other variables must have made a difference–in other words **there must have been differences in at least one of distance from the freeway, temperature or rain status between the two locations at the times the measurements were taken.** If the average distance, temperature or rain status for the measurements at the two locations had been the same then adjusting for these factors wouldn't have mattered–in essence we would have been plugging in the same values for the variables at both locations and the difference wouldn't have changed. Now logically at any given school on any given day, the distance from the freeway is probably very similar at the parking lot and the play yard and it is presumably either raining at that school or not. Thus unless Dr. Green did yard measurements at one set of schools and parking lot measurements at another lot of schools, or did parking lot measurements on rainy days and yard measurements on sunny days, these variables shouldn't be different between the locations. However we were told in the problem statement that she took measurements in the parking lot in the morning when people were arriving and measurements in the yard during afternoon recess when the children were there. You would expect it on average to be cooler in the morning and warmer in the afternoon. The multiple regression model tells us that on average pollution is higher when the temperature is higher. Thus measuring the parking lot in the morning when it was cooler probably made it look relatvely better compared to the play yard which was measured in the afternoon when it was warmer. In Question 3, we saw that pollution in the parking lot **in the morning** was similar to pollution in the yard **in the afternoon**. However if we adjust for the difference in temperature we find that **under comparable conditions** the pollution level is worse in the parking lot than the yard which is much more what we would have expected. Most people managed to get at the idea that in the MLR we were adjusting for other variables which must somehow have made a difference but few people got at the critical point that there must have been a difference between the locations on the three additional variables, and even fewer came up with what that difference might have been that would account for the particular change in results that we observed. I try very hard to make things in these problems work out as you would expect from common sense and I rarely provide information (e.g. that measurements were taken in the parking lot in the morning and in the yard in the afternoon) unless there is some point to them. The moral is, trust your intuition–and if it seems like I have provided extraneous information, look as you go through the

problem to see if it could have any bearing on the answers :)