Biostatistics 201A, 2011 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	76.06
Median	77.75
Standard Deviation	12.92
Maximum	96.00
Minimum	35.00
1st Quartile	68.00
3rd Quartile	87.00

Approximate Grade	Score Range	Number In Range
A	87+	13
A-	80-86.5	8
B+	75-79.5	8
B	60-74.5	16
B-	50-59.5	5
Below	< 50	1

THE STORY BEHIND THE EXAM: Professor Fatima "Fati" Assad, a nutrition researcher at my favorite school, the University of Calculationally Literate Adults, is interested in dietary and genetic influences on blood cholesterol levels since high cholesterol is a significant risk factor for heart disease. She is also in the process of developing a new treatment for high cholesterol and wants to know on whom it will be most effective. During the exam you will help her analyze some of her preliminary data and interpret the results.

Question 1: Garbage In, Garbage Out? (18 points, 15 minutes)

Dr. Assad believes that one major contributor to high cholesterol levels is poor diet. She has collected data on n = 62 people and recorded, among other things, each subject's blood cholesterol level, Y, in mg/deciliter, and their average daily caloric intake, X. A STATA printout for a simple linear regression using these two variables is shown below along with some follow-up analyses. Use the output to answer the questions on the subsequent pages. You may find it useful to know that total normal levels of cholesterol should be below 200 while levels from 200-239 are borderline high and levels above 240 are considered high. The average cholesterol level in Dr. Assad's sample was $\bar{Y} = 220$ and the average caloric intake was $\bar{X} = 2000$ calories per day, a typical healthy adult diet.

```
. regress cholesterol calories

      Source |       SS       df       MS              Number of obs =      62
-------------+------------------------------           F(  1,    60) =   29.64
       Model |    32269     1       32269              Prob > F      =  0.0000
    Residual |    65331    60        1089              R-squared     =    0.33
-------------+------------------------------           Adj R-squared =    0.32
       Total |    97600    61        1600              Root MSE      =    32.0

------------------------------------------------------------------------------
 cholesterol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    calories |     0.098       0.018     5.44   0.000       0.062       0.134
       _cons |    23.671      36.306     0.65   0.517     -48.954      96.296
******************************************************************************
. adjust calories = 1800, se ci
-------------------------------------------------------------------------------
     Dependent variable: cholesterol   Covariate set to value: calories = 1800
-------------------------------------------------------------
      All |         xb        stdp          lb          ub
----------+--------------------------------------------------
          |    *******   (5.52887)    [189.308    211.426]
-------------------------------------------------------------
Key: xb = Linear Prediction  stdp = Standard Error [lb, ub] = [95% Confidence Interval]
******************************************************************************
. adjust calories = 1800, stdf ci
-------------------------------------------------------------------------------
     Dependent variable: cholesterol  Covariate set to value: calories = 1800
-------------------------------------------------------------
      All |         xb        stdf          lb          ub
----------+--------------------------------------------------
          |   ********   (33.4577)    [133.442    267.293]
-------------------------------------------------------------
Key: xb = Linear Prediction  stdf = Standard Error [lb, ub] = [95% Prediction Interval]
```

Part a (3 points)

Give a brief interpretation of R^2 and R^2_{adj} and explain whether by this measure caloric intake is useful for explaining cholesterol levels.

Solution: R^2 and R^2_{adj} give the percentage of variability in Y that is explained by the predictor variables, X. R^2_{adj} takes into account the degrees of freedom to provide an unbiased estimate and penalizes you for including too many predictors that do not add significant explanatory power. In multiple regression it is a good idea to always use R^2_{adj} though in simple linear regression, as here, there is usually not much difference. In this problem R^2 and R^2_{adj} tell us that 32%-33% of the variability in peoples' cholesterol levels can be explained by how many calories they consume per day. This is a relatively low percentage so the model is not doing an outstanding job. However this is not really surprising since there are many factors that affect cholesterol levels and in particular the kind of food being consumed rather than just the simple calorie count is probably important. In this sense you could argue that for a single variable we are not doing such a bad job.

Part b (5 points)

Give a brief interpretation of RMSE and discuss whether by this measure caloric intake is useful for explaining cholesterol levels, quoting appropriate numbers to justify your answer.

Solution: The RMSE is, roughly, the average error we make when using the X variables to predict Y in our sample. To judge whether or not the errors are big we need a reference point. That can either be the values in the sample ($\bar{Y}$ and the minimum or maximum values of Y are common choices) or some idea what meaningful values are in the context of the problem. Here $RMSE = 32$ meaning that when we use daily caloric intake to predict cholesterol level we are typically off by 32 mg/dL. We note that in our sample the average cholesterol level was 220 mg/dL and that moreover the threshold for having high cholesterol os around 200 mg/dL. Based on these numbers we are typically making around a 15% error (32/220). This is pretty good but not absolutely spectacular.

Part c (7 points)

Lily Lipid is on an 1800 calorie per day diet and has a blood cholesterol level of 300. Is this unusual? Discuss this (i) by using Dr. Assad's regression model to predict the typical cholesterol level of a person with this diet (which seems strangely to have disappeared from the printouts!) and (ii) by using an appropriate interval for that prediction. Your answer should make it clear what interval you are using for part (ii) and why.

Solution: First we need to plug into the regression equation to find the value the model would predict for Lily Lipid. We have

$$\hat{Y} = b_0 + b_1 X = 23.67 + .098(1800) = 200.07$$

This is WAY below Lily's actual value of 300. Our model has not made a very good prediction of her cholesterol level. Moreoverm looking at the the printout we see that 300 is well beyond the upper end of either the confidence interval for the average cholesterol level of people on an 1800 calorie diet ([189.308, 211.426]) or the prediction interval for an individual person with an 1800

calorie per day diet ([133.442 267.293]). Here since we are specifically talking about whether **Lily's** cholesterol level is unusual we should be using the prediction interval. The model suggests that the percentage of individuals with the same dietary intake as Lily's who have a cholesterol level as high as 300 mg/dL is very small. Her situation IS unusual.

Part d (3 points)

Give one statistical and one real-world possibility for what could have caused the result you saw in part (c).

Solution: Make sure you understand the distinction between a "statistical" reason and a "real-world" reason. From a statistical point of view we could simply have been unlucky, either in getting an unusual sample which gave us poor estimates of the model parameters (and hence poor predictions) or in having chosen an individual, Lily, who was at the extreme end of the distribution for her diet. After all, the 95% prediction interval is supposed to include the cholesterol levels for most people with 1800 calorie per day diets but not all of them–Lily could simply have been one of those not covered by the interval. From a practical point of view it seems more likely that there is something unusual about Lily. For example, she could be getting most of her calories from fat so that even though the overal count isn't so high, her cholesterol intake is. She could be dieting now but have built up high cholesterol levels from her previous diet. Or, she could have a genetic tendency towards high cholesterol levels (see Questions 2 and 3) in which case even a good diet might not be enough to keep her cholesterol levels normal.

Question 2: Fat City (42 points, 55 minutes)

In addition to cholesterol level, Y, and caloric intake, X_1, Dr. Assad also collected some other data on the subjects from Question 1. In particular, she recorded average daily dietary fat intake, X_2 (in grams per day), age, X_3 (in years) and gender ($X_4 = 1$ for males and $X_4 = 0$ for females). She also obtained data on a particular mutation in the LDLR (low-density lipid receptor) gene which is associated with familial hypercholesterolemia (FH), a genetic disorder characterized by extremely high levels of LDL or "bad" cholesterol. A person can have two normal copies of the gene, one normal and one abnormal copy of the gene (heterozygote), or two abnormal copies (homozygote). This latter genotype is rare and very serious. Dr. Assad has included two indicator variables in her model to code genetic status: $X_5 = 1$ if the subject is a heterozygote and 0 otherwise and $X_6 = 1$ if the subject is a homozygote and 0 otherwise. Some summary statistics for these variables are given below and the corresponding multiple regression printout is shown on the next page.

```
. cor cholesterol calories fat age
(obs=62)
             | choles~l calories      fat      age
-------------+------------------------------------
 cholesterol |   1.0000
    calories |   0.5750   1.0000
         fat |   0.6879   0.8511   1.0000
         age |  -0.0210  -0.2328  -0.1791   1.0000

*******************************************************************************

. ttest cholesterol, by(gender)
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.6159         Pr(|T| > |t|) = 0.7682          Pr(T > t) = 0.3841

*******************************************************************************

. ttest cholesterol, by(heterozygote)
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7304         Pr(|T| > |t|) = 0.5391          Pr(T > t) = 0.2696

*******************************************************************************

. ttest cholesterol, by(homozygote)
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000

*******************************************************************************
```

```
. regress cholesterol calories fat age gender homozygote heterozygote

      Source |       SS       df       MS              Number of obs =      62
-------------+------------------------------           F(  6,    55) =  144.17
       Model |   91765.4     6   15294.24              Prob > F      =  0.0000
    Residual |    5834.6    55     106.08              R-squared     =  0.9402
-------------+------------------------------           Adj R-squared =  0.9337
       Total |     97600    61       1600              Root MSE      =    10.3

------------------------------------------------------------------------------
 cholesterol |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    calories |   .0125392   .0108898     1.15   0.255    -.0092845    .0343629
         fat |   1.141492   .1265784     9.02   0.000     .8878234    1.395161
         age |   .4878673   .0914532     5.33   0.000      .304591    .6711435
      gender |  -5.189587   2.624269    -1.98   0.053    -10.44874    .0695658
heterozygote |   21.54936   2.935958     7.34   0.000     15.66557    27.43315
  homozygote |   74.91659   3.679003    20.36   0.000      67.5437    82.28947
       _cons |   81.44491   17.73894     4.59   0.000     45.89527    116.9945
------------------------------------------------------------------------------
*******************************************************************************
. test heterozygote = homozygote

 ( 1)  heterozygote - homozygote = 0

       F(  1,    55) =  213.85
            Prob > F =    0.0000
*******************************************************************************
```

Part a (5 points)

Based on their **individual** relationships with cholesterol level which of these variables would you expect to be good predictors? Briefly justify your answers.

Solution: For the categorical variables the t-tests tell us whether there is an individual relationship with cholesterol while for the continuous predictors we need to look at the correlations with cholesterol. We note from the t-tests that there is a significant difference (p-value less than .05) between homozygotes and non-homozygotes for the LDLR gene but not between men and women or between the heterozygotes and non-heterozygotes. Note that this last comparison is not a great one since the non-heterozygote group includes both the people with two normal copies of the gene and the homozygotes–so we don't have a clear idea of how the heterzygotes compare to each of the other two genotype groups. We therefore expect that the indicator for homozygous status will be a useful predictor but gender will not and we are uncertain about the indicator for the heterozygous group. Looking at the correlation table we see that both calories (r = .575) and dietary fat (r = .688) have quite strong correlations with cholesterol so we would expect these to be good predictors.

However age does not have a strong individual relationship with cholesterol (r = -.02) and therefore may not be a good predictor. Of course all of these results are subject to change in a multiple regression model if there is any confounding or multicollinearity going on!

Part b (6 points)

Give units and real-world interpretations for b_2 and b_4, the coefficients of the dietary fat and gender variables. Make sure your answers incorporate the numerical values of the coefficients.

Solution: The most common mistakes on this problem were to forget to include the units or numeric values of the coefficients, to neglect to mention that all the other variables had to be held fixed, or to not give the (intuitive) direction of the relationship between the predictor variables and the outcome. First, for dietary fat, $b_2 = 1.14$ tells us that **all else equal** or **assuming two people with the same gender, genotype, caloric intake and age** on average each additional gram of fat in a person's diet is associated with a 1.14 mg/dL **higher** cholesterol level. For gender, $X_4 = 1$ for males so the coefficient value $b_4 = -5.19$ tells us that a man will have a cholesterol level that is 5.19 mg/dL **lower** than an **otherwise equivalent** woman.

Part c (6 points)

Is caloric intake a significant predictor in the multiple regression model at $\alpha = .05$? Does this result differ from what you saw in Question 1 or in part (a) above? Explain what you think might have happened, providing any available evidence to support your theory.

Solution: From the printout, the p-value for the t-test for the coefficient of the calorie variable is .255, definitely higher than our usual significance level of $\alpha = .05$. Caloric intake is therefore not a significant predictor as part of this mode, even though we saw above that it had a high correlation with cholesterol level and indeed in Question 1 we found that by itself it did have a significant relationship with cholesterol. The key to what has happened lies in the interpretation of the t-tests in a multiple regression. We are testing whether caloric intake explains a significant amount of **additional** variation in cholesterol levels **after accounting for fat intake, gender, age and genotype.** It appears that although caloric intake is related to cholesterol levels that relationship is explained by one or more of these other factors. Intuitively it seems likely that it is fat intake that is accounting for the relationship. Dietary fat probably has a more direct link to cholesterol levels but of course high fat intake results in a high calorie diet. Our intuition is confirmed by the very high correlation (r = .85) between caloric intake and dietary fat. This is of course an example of multicollinearity, our current study topic. Have a think about whether you feel this is a case of simple confounding or potentially one of mediation....Note also that you did not have to write out all the details of the hypothesis test here (which many people did). Unless I ask for more simply citing the p-value for a test is sufficient and will save you a lot of time!

Part d (5 points)

Are there significant differences in cholesterol level by genotype after adjusting for diet, age and gender? Carefully justify your answers using $\alpha = .05$. (You do not need to write out all the details

of the tests.)

Solution: To answer this question we first need to look at the p-values for β_5 and β_6, the coefficients of the two genotype indicators. This will tell us how people who are heterozygotes and homozygotes for the LDLR genotype compare to the people with two normal copies of the gene. Both these p-values are essentially 0 so we can be quite sure that both heterozygotes and homozygotes have significantly different cholesterol levels from people with two normal copies. Second, to compare the heterozygotes to the homozygotes we need to look at the follow-up F test given after the regression printout. The p-value for this test was also 0, indicating that heterozygotes and homozygotes have different cholesterol levels, even after adjusting for all the other variables in the model.

Part e (5 points)

Dr. Assad's graduate student says she has a theory that women have higher cholesterol levels than men (after adjusting for age, genotype and diet). Write down the mathematical hypotheses and p-value associated the test she wishes to perform and give your real-world conclusions using $\alpha = .05$.

Solution: Since Dr. Assad's graduate student wants to specifically prove that women have **higher** cholesterol levels than men rather than just checking whether there is a gender **difference** this is a **one-sided** test. In the problem statement we note that $X_4 = 1$ for men so the sign of the regression coefficient will indicate how men compare to women. Trying to show that women have higher cholesterol levels than men is equivalent to showing that men have **lower** cholesterol levels than women which corresponds to a **negative** coefficient for X_4. Our hypotheses are therefore

$H_0 : \beta_4 \geq 0$: Men have the same or higher cholesterol levels than women–or equivalently women have the same or lower cholesterol levels than men.
$H_A : \beta_4 < 0$: Men have lower cholesterol levels than women-or equivalently women have higher cholesterol levels than men.

Our test statistic is $t_{obs} = -1.98$ and the correpsonding p-value is $P(t_{55} \leq -1.98) = .053/2 = .0265$. Note that to get the one-sided p-value we had to divide the STATA p-value in half since STATA always gives us the p-value for the 2-sided test. This p-value is smaller than $\alpha = .05$ so we reject the null hypothesis and conclude that women have higher cholesterol levels than men in accordance with the graduate student's theory.

Part f (3 points)

The graduate student from part (e) admits that she developed her theory after noticing that the women in this sample had higher average cholesterol levels than the men. Does this affect how you view your conclusions from part (e)? Discuss briefly.

Solution: This does change our answer to part (e). If the student didn't hypothesize ahead of time (without looking at the data) that women had higher cholesterol levels then she is really doing a 2-sided test. Effectively she used the data to eliminate one of the possible alternatives and she has to count that look at the data when determining her p-value. (Note that otherwise no one would ever do a 2-sided test–they would just look at their data, pick whichever direction the data

favored, and reject much more often–and as a result they would make many more type I errors. Even when nothing is going on the sample data will always tilt slightly in one direction or the other by chance and if you look at the data to choose your hypothesis you will be assuming those chance leanings are real!) However, if the student does a 2-sided test her p-value is .053, not .0265, and at significance level $\alpha = .05$ she will fail to reject the null hypothesis–she will not be able to conclude that there is a gender difference in cholesterol levels after accounting for caloric intake, age and genotype.

Part g (7 points)

According to your best estimate from this model, how much would you need to reduce your daily dietary fat intake to lower your cholesterol by 10mg/dL? How much would you need to reduce it by to be 95% sure of lowering your cholesterol by 10 mg/dL? What assumption are you inherently making when you do this calculation?

Solution: People always seem to have trouble with questions that involve manipulating the individual regression coefficients but the coefficients are the key to understanding what a regression model is telling you so this is an important thing to practice. In part (b) we saw that for every extra gram of dietary fat consumed per day on average cholesterol level was 1.14 mg/dL higher. Equivalently for every 1 gram less of fat consumed, cholesterol levels would be on average 1.14 mg/dL lower. Thus eating 2 grams less of fat would be associated with 2.28 mg/dL less cholesterol and so on. To find out how many grams less we need to consume to be 10 mg/dL lower in cholesterol level we need to solve $10 = 1.14$x. We find that we need to consume $10/1.14 = 8.77$ fewer grams of fat per day. This is our best estimate. However if we want to be 95% sure we need to look at the confidence interval. According to the low end of the confidence interval it is possible that cholesterol levels drop only .888 mg/dL for each gram of fat elimiated from the diet. Thus we may have to consume $10/.888 = 11.26$ fewer grams of fat to get the desired reduction in cholesterol. Note that I used the lower end of the confidence interval because this is the worst case scenario for cholesterol reduction–it gives me the smallest reduction in cholesterol I would expect to get for each gram of fat elimiated. Of course all of this is assuming that the relationship between dietary fat and cholesterol level is **causal** so that when I change my fat intake it will result in my cholesterol level going down. In practical terms this assumption doesn't seem too unreasonable but of course correlation is not causation. In particular it turns out that for people with familial hypercholesterolemia simply changing the diet as very little effect on cholesterol levels.

Part h (5 points)

Use the model to estimate the average cholesterol level of 35 year old women whose typical dietary intake is 1800 calories and 50 grams of fat per day and who are homozygous for the LDLR genetic mutation.

Solution: All we have to do is plug into the regression equation noting that for a woman the gender indicator is $X_4 = 0$ and that for a homozygote the the genotype indicators are $X_5 = 0, X_6 = 1$. The resulting prediction is

$$\hat{Y} = 81.445 + +1.141 * 50 + .012 * 1800 + .488 * 35 - 5.190 * 0 + 21.549 * 0 + 74.917 * 1 = 252.092$$

Our prediction is that this person will have a cholesterol level of 252.092 mg/dL. Note that this is a possible explanation for the extremely high cholesterol level of Lily Lipid in problem 1. Taking into account fat intake, age and genotype we got a much higher predicted cholesterol level than we did from the model based on just caloric intake.

Part i (Optional Bonus)

Suppose in part (e) Dr. Assad's graduate student had been trying to show that men had higher cholesterol than women. What would her p-value have been? Explain briefly.

Solution: The two 1-sided tests are complementary so their p-values have to add up to 1. Specifically, the p-value in part (e) when we were trying to prove $H_A : \beta_4 < 0$ was $P(t_{55} \leq -1.98)$. If we are trying to prove the opposite, $H_A : \beta_4 > 0$ then our p-value will be $P(t_{55} \geq -1.98) = 1 - .0265 = .9735$. The inequality in the p-value calculation for a 1-sided test always matches the inequality in the statement of the alternative hypothesis since the p-value is calculating how likely you are to see something as or more favorable to the alternative hypothesis as your data assuming the null is true. In this case since the data do support the idea that women have higher cholesterol levels (or men have lower ones) they can not possibly support the idea that men have higher cholesterol levels and so the p-value is huge!

Part j (Optional Bonus)

Suppose instead of using indicator variables for the genotype groups Dr. Assad had let X_5 be the number of abnormal copies of the LDLR gene the person had (with possible values 0, 1 or 2). What would that have assumed about the differences in cholesterol levels between the genotype groups and roughly what do you think the estimated coefficient of the new X_5 would have been? Explain briefly.

Solution: If we treat the number of genetics mutations as a continuous variable our interpretation of the coefficient is the change in cholesterol level associated with each ADDITIONAL abnormal copy of the gene. This means we are assuming that the difference between having no abnormal copies and one abnormal copy is the same as the difference between having one abnormal copy and two abnormal copies. Note that we suspect this isn't true since the jump from two normal copies to being a heterozygote was only about a third of the jump from having two normal copies to being a homozygote. Since that latter jump, from 0 to 2 abnormal copies, was about 75 mg/dL you might think that the per abnormal copy jump would be around 37.5 mg/dL. However we also have an estimate that going from no abnormal copies to being a heterozygote is associated with an additional 21.5 mg/dL of cholesterol. Our esimtate will be between these two values and give that there are probably many fewer homozygous subjects in the data set than heterozygotes the estimated value will probably be closer to 21.5 than 37.5.

Question 3: Pick Six (40 points, 50 minutes)

Next Dr. Assad tries out her new cholesterol lowering treatment. Based on the results of Question 2 she has decided to randomly assign her new treatment, T, or a placebo, P to subjects in each of the three genotype groups (normal=NN, heterozygous=HN, homozygous=HH), giving her k=6 groups. An ANOVA printout for her data is shown below. Note that the table of pairwise mean differences and p-values has NOT been adjusted for multiple comparisons.

```
. oneway cholesterol group, tabulate
           |       Summary of Cholesterol
     group |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
      NN+P |         200        19.5          32
      HN+P |         220        19.5          32
      HH+P |         300        24.0           8
      NN+T |         180        19.5          32
      HN+T |         190        19.5          32
      HH+T |         285        24.0           8
-----------+------------------------------------
     Total |         208        38.5         144
                       Analysis of Variance
    Source              SS         df       MS           F     Prob > F
------------------------------------------------------------------------
Between groups      157256          5    31451       78.63     0.0000
 Within groups       55200        138      400
------------------------------------------------------------------------
    Total           212456         143    1486
**************************************************************************
          Comparison of Cholesterol by Group--NOT corrected for Multiple Comparisons
Row Mean-|
Col Mean |       NN+P       HN+P       HH+P       NN+T       HN+T
---------+-------------------------------------------------------
    HN+P |         20
         |     0.0001
         |
    HH+P |        100         80
         |     0.0000     0.0000
         |
    NN+T |        -20        -40       -120
         |     0.0001     0.0000     0.0000
         |
    HN+T |        -10        -30       -110         10
         |     0.0475     0.0000     0.0000     0.0475
         |
    HH+T |         85         65        -15        105         95
         |     0.0000     0.0000     0.1359     0.0000     0.0000
```

Part a (5 points)

According to the printout, which pairs of groups do **not** have significantly different cholesterol levels **without** adjusting for multiple comaprisons? Use this information to describe as precisely as you can the treatment effects and genotype differences.

Solution: All of the p-values in the above table for pairwise comparisons are less than $\alpha = .05$ except the one for the test comparing the homozygous treatment and control groups (HH+T and HH+P, p-value = .1359). Thus it appears that this is the only pair of groups where we don't have a significant difference in cholesterol levels. Up to this point everyone was fine. However there were a lot of problems with the subsequent interpretations. I asked you to discuss (i) the treatment effects and (ii) the genotype differences based on the above data. The treatment effects refer to the difference between the treatment group and the placebo group within each genotype. Based on the above comparisons we have evidence of a treatment effect for the NN and HN genotypes but not for the the HH genotype. Note that the tests just tell us there is a difference. We can tell the treatment effect is a positive one by looking at the group means in the first part of the printout and seeing that in each genotype the cholesterol level is lower in the treatment group than the control group–it just isn't enough lower in the HH treatment group to be significant. I also asked you to comment on the genotype differences. Within each treatment arm of the study we see that the cholesterol levels are lowest in the NN group, intermediate in the HN group and highest in the HH group and all of those comparisons are significant. Thus we clearly have evidence of genotype differences.

Part b (4 points)

Now suppose you wanted to apply the Bonferroni correction for multiple comparisons to the set of tests from part (a). What would you use as your significance level, α^*, for the individual tests? How would this change your answers to part (a) if at all?

Solution: The Bonferroni approach to multiple comparisons says that if you overall want to have a significance level of α for a set of m tests then for each individual test you should use a significance level of $\alpha^* = \alpha/m$. Here we have $\alpha = .05$ and we have are doing a total of $m = 15$ tests. Note that to find m we can either just count the tests on the printout or we can note that there are k=6 groups and we want all possible pairs out of those 6 groups or "k choose 2" which is (6)5)/2 =15. Thus we need to use $\alpha^* = .05/15 = .0033$ for the individual tests. At this level two additional tests will become insignificant–the comparison of the heterozygous treatment group (HN+T) to the normal placebo group (NN+P) and the two-normal copies treatment group (NN+T). The implication is that after treatment the heterozygotes may be back to the level of a normal untreated group or even to the level of a normal treated group.

Part c (7 points)

Find the effect sizes (Cohen's d values) for comparing treatment to control subjects for each of the three genotype groups (e.g. NN+T vs NN+P) and explain what this tells you about the relative effectiveness of the intervention for the different groups of people. (Note: You have two choices as to what to use for your standard deviation estimate as part of the effect size calculations. Briefly

explain your reasoning for picking whichever one you use.)

Solution: The effect size for comparing the difference in means between two groups is

$$d = \frac{\bar{Y}_1 - \bar{Y}_2}{s}$$

where s is some sort of pooled estimate of the standard deviation (assumed to be) shared by the two groups. In this problem we have two reasonable choices for that pooled estimate of the standard deviation. One is to calculate it based on the two groups being compared; the other is to use $\sqrt{MSW}$, the estimate of the common standard deviation shared by all 6 groups. After all, in an ANOVA context we generally assume all the groups have the same variability. If that is true then $\sqrt{MSW}$ is the most accurate estimate because it uses all the available information from a much larger sample. Here the group standard deviations are 19.5 for the NN and HN groups and 24 for the two HH groups so the assumption doesn't seem too unreasonable and I would base my effect sizes on $\sqrt{MSW} = \sqrt{400} = 20$. Many people just used 19.5 for the effect sizes in the NN and HN groups and 24 for the HH group–we accepted this if you made the argument that it looked like the variability was really higher in the HH group and so it was potentially more accurate to use different values. However you needed to make a case for it. Some people tried to use the value 38.5 from the "total" row of the summary table. This value corresponds to the overall standard deviation in cholesterol levels **ignoring** group membership and is not thje appropriate measure here. We need to be using a measure of the standard deviation within each of the groups. Using $\sqrt{MSW} = 20$ the effect sizes were particularly easy to calculate:

$d_{NN} = \frac{200-180}{20} = 1$
$d_{HN} = \frac{220-190}{20} = 1.5$
$d_{HH} = \frac{300-285}{20} = .75$

Using the conventions of Cohen which we learned in class a standard medium effect is around $d = .5$ and a large effect is $d = .8$ so these all seem like quite large effect sizes though of course to answer that most precisely we should actually ask what would be clinically meaningful in terms of cholesterol levels. I actually asked you to comment on the **relative** effectiveness of the treatments. Based on the effect sizes it seems the treatment is most effective in the heterozygous group, followed by the normal group and least effective in the homozygous group. This is consistent with what we saw in parts (a) and (b).

Part d (9 points)

Dr. Assad's graduate student did some power calculations prior to the treatment study. Specifically, she found the minimum detectable effect size with n=32 subjects per group for a two-sample t-test using the standard settings of 80% power and a significance level of $\alpha = .05$.

```
. sampsi 0 .7, sd(1) n(32)

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 =       .7
           sd1 =        1
           sd2 =        1
sample size n1 =       32
            n2 =       32
         n2/n1 =     1.00

Estimated power:
         power =   0.7996
```

(i) What did she find as the minimum detectable effect size and what does it say about how well powered the study was?
(ii) Was her choice of $\alpha = .05$ a good one given the analyses that were subsequently performed? If so why? If not, what would have happened to the power if she had used a more appropriate α?
(iii) The available sample size for the homozygous groups (HH+T and HH+P) was only n=8 each. Explain briefly the effect this would have on the power calculation for that comparison.

Solution: (i) The first part of this question in particular created a lot of problems. From the STATA printout we are calculating the power for comparing groups with means 0 and .7 and standard deviations of 1 leading to an effect size of

$$d = \frac{.7 - 0}{1} = .7$$

According to the printout with 32 subjects per group we have approximately 80% power to detect an effect of this size using a significance level of $\alpha = .05$. Since 80% is generally considered the acceptible power cutoff $d = .7$ is the **smallest** effect size we can detect with reasonable power. However $d = .7$ is a fairly large effect size. This means our study is only powered to detect large effects which is not so good. If the real effect is only small or medium we will not have a good chance of seeing it. Make sure you understand the logic. In general having a large **actual** effect size is good–your power to detect large effects is higher, all else equal. However having the **minimum detectable effect** be big is bad–it means that you can only detect large effects. Note also that we probably do not know ahead of time that our effect size will be big–in particular we don't have values from part (c) until after we have done our study–so we can't use them to say that we would have been happy about our power ahead of time. As it turned out our effects were large so we did have enough power for most of the comparisons but unless we could have made an argument in advance that our effect sizes were likely to be large a funding agency would not have been happy with these numbers. A few people made an argument to the effect that only large effect sizes would have amounted to clinically meaningful changes in cholesterol levels and that therefore the study was actually adequately powered to detect any treatment effects that would be important. That would be a reasonable argument although note that it does require you to have an idea in advance

of what the standard deviation is.

(ii) The graduate student's choice of $\alpha = .05$ as the basis for her power calculations is **not** appropriate. As we saw in parts (a) and (b) she and Dr. Assad planning to do all the pairwise comparisons of the group means. If you are performing multiple tests you need to use some sort of correction, e.g. Bonferroni, for the individual tests. Since the student is presenting power for an individual t-test she needs to do the calculation using the significance level for that individual test. If the student had used $\alpha^* = .0033$ as in our calculations from part (b) she would have made her power much **lower**. A smaller significance level makes it much harder to reject the null hypothesis which makes power–the probability of finding effects that are really there–much smaller. Note: A number of people tried to tell us here that the standard deviation gets bigger when n is smaller nd hence the power is lower. The first part of that statement is NOT true. The standard deviation of the individual data points does not depend on the sample size. It is the **standard error**–that is the standard deviation of the sample means–that gets smaller as the sample size gets larger–and thereby affects the power.

(iii) The sample sizes for the homozygous groups are much smaller (n=8) than the ones used in the power calculation (n=32) which will again make the power **lower** than what is shown on the graduate student's printout. The smaller your sample size the less information you have and the less chance you will have convincing evidence of an effect.

Part e (5 points)

The printout below shows an hypothesis test for a particular linear combination. Explain intuitively what Dr. Assad is trying to test and what her conclusion will be.

```
. test (group.NNT + group.HNT + group.HHT)/3 = (group.NNP + group.HNP+group.HHP)/3

.33group.NNT + .33group.HNT + .33group.HHT - .33group.NNP - .33group.HNP - .33group.HHP = 0

       F(  1,   138) =   28.17
            Prob > F =    0.0000
```

Solution On one side of the test is the average cholesterol level of the three groups that received Dr. Assad's treatment and on the other side of the test is the average of the three groups that did not receive treatment. Thus Dr. Assad is testing whether or not there is an **overall** treatment effect, ignoring genotype. The p-value for her test is 0, meaning she can reject the null hypothesis of no difference between the treatment and placebo groups and conclude that there is a treatment effect–the groups are **different.** Note that Dr. Assad did a two-sided test. She did not specifically try to show that the treatment group was **better** than the placebo group–just that there was a difference. Of course, by looking at the group means we can tell that the treatment effect was in the direction of lowering cholesterol but that is not what we have formally tested.

Part f (10 points)

Dr. Assad believes that her treatment will have a greater effect on cholesterol levels in heterozygous (HN) subjects than it will in homozygous (HH) subjects.

(i) Write down the linear combination she is interested in and give your best estimate of it based on the data at the beginning of the problem.
(ii) Calculate the standard error associated with the linear combination Dr. Assad wishes to test.
(iii) Compute a 95% confidence interval for the linear combination and use it to evaluate Dr. Assad's theory. (Note–you may not be able to get the exact value of t that you need for the confidence interval. Just use a reasonable approximation.)

Note: If you can't figure out which linear combination Dr. Assad wants you can still get partial credit by showing how you would do the problem for an LC you do understand.

Solution: We gave up to half credit if you gave a completely correct solution for an aribtrary linear combination.

(i) The treatment effect for a given genotype is simply the difference between the treated and untreated groups with that genotype. Thus the treatment effect for the heterozygous subjects is $\mu_{HNP} - \mu_{HNT}$ and similarly the treatment effect for the homozygous subjects is $\mu_{HHP} - \mu_{HHT}$. Dr. Assad wants to prove that

$$\mu_{HNP} - \mu_{HNT} > \mu_{HHP} - \mu_{HHT}$$

Moving all the pieces to one side of the equation we see that her linear combination of interest is

$$LC = (\mu_{HNP} - \mu_{HNT}) - (\mu_{HHP} - \mu_{HHT})$$

Note that although we are trying to show $LC > 0$ the "greater than 0" piece is NOT part of the linear combination. The linear combination is just the expression involving the means. Also notice that we are not talking about any sort of averaging here. We want the difference between the treatment effect for heterozygotes and the treatment effect for homozygotes so we do NOT need to divide everything by 2! Our best estimate for the linear combination is obtained by plugging in the sanple group means. We have

$$\hat{LC} = (220 - 190) - (300 - 285) = 30 - 15 = 15$$

(ii) The standard error associated with the linear combination $LC = \sum c_j \mu_j$ is just

$$\sqrt{MSW \sum_{j=1}^{k} \frac{c_j^2}{n_j}}$$

Here MSW = 400, the constants in our linear combination are all ± 1 and the group sizes are 32 for the heterozygous groups and 8 for the homozygous groups so the resulting standard error is

$$s.e.(\hat{LC}) = \sqrt{400(\frac{1^2}{32} + \frac{(-1)^2}{32} + \frac{1^2}{8} + \frac{(-1)^2}{8})} = 11.18$$

(iii) For a confidence interval for a linear combination we use the formula

$$\hat{LC} \pm t_{alpha/2,n-k} s.e.(\hat{LC})$$

We got the estimated linear combination and standard error in parts (i) and (ii) so we just need the appropriate t-value. We want a 95% interval so $\alpha = .05$ and $\alpha/2 = .025$. In the ANOVA we had n = 144 and k = 6 groups so our degrees of freedom are 138. This value doesn't appear on the t-table. The closest value is for 120 degrees of freedom would is $t_{.025,120} = 1.98$. This is more conservative than using 1.96, the value associated with an infinite number of degrees of freedom. (The larger t value makes the interval wider which means it is moer likely to include the true LC value.) Many people calculated their degrees of freedom just based on the groups involved in the linear combination. Remember that the degrees of freedom come from your estimate of the standard deviation which here is MSW. The resulting interval is

$$15 \pm (1.98)(11.18)$$

or $[-7.14, 37.14]$. This means that the treatment effect for the heterozygotes could be anywhere from 7 mg/dL less than for homozyhotes to 37 mg/dL greater than for homozygotes. Since the interval includes values corresponding to either group having a bigger treatment effect (or no difference in treatment effect if LC = 0) Dr. Assad does not have enough evidence to conclude that the treatment effect is bigger for heterozygotes than homozygotes.

One additional note. Dr. Assad is in fact interested in a 1-sided test so you could make an argument that if she wants to be 95% sure she really only needs a 90% confidence interval. However I explicitly asked you to compute a 95% interval and use it to evaluate her theory....

Part g (Optional Bonus):

Note that the p-value for comparing the treatment and control subjects with the homozygous genotype (HH+T to HH+P) is much larger than that for comparing the normal treatment group (NN+T) to the heterozygous treatment group (HN+T), even though the mean difference between the two homozygous groups is larger than that for the difference between the other two treatment groups. Explain briefly how this could happen.

Solution: First note that I've asked you to compare the mean difference of HHT and HHP (which is 15) to the mean difference of HNT and NNT (which is 10). I have not said the treatment effect for the HH group was bigger than the treatment effect for either the HN or NN groups (which it is not). Several people got confused by this. The test depends on the mean difference, the pooled estimate of the standard deviation, $\sqrt{MSW}$, and the sample sizes. Since we are using the same $\sqrt{MSW}$ for all the tests and the mean difference is bigger for the HHT for HHP comparison the only way it could have come out less significant is because of the sample sizes–and indeed the sample sizes for the HH groups at n=8 are much smaller than the n=32 that we have for the HNT and NNT groups. The smaller sample size results in a much bigger standard error and hence a less significant result. A number of people tried to tell us here (and in earlier parts of this problem) that the standard deviation gets bigger when n is smaller. That is NOT true. The standard deviation of the individual data points does not depend on the sample size. It is the **standard error**–that is the standard deviation of the sample means–that gets smaller as the sample size gets larger.