November 28th, 2011

Final Exam Practice Problems With Solutions

Logistic Regression Practice

(1) Logistic Regression Basics:

(a) Explain what the response variable is in a logistic regression and the tricks we use to convert this into a mathematical regression equation.

Solution: In a logistic regression the response variable, Y, is an indicator saying whether or not you have a particular characteristic, say lung cancer. The problem is that the value of an indicator is always 1 or 0–this is how we turn something qualitative into something quantitative. Unfortunately a model of the form $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ doesn't produce values that are exactly 0 or 1. Thus instead we focus on modeling the **probability** that Y=0 or Y=1. Even this isn't quite good enough as a probability is between 0 and 1 and their is no certainty that $\beta_0 + \beta_1 X_1 + \beta_2 X_2$ will lie in that range. Thus we do one more trick which is to take the odss that Y=1, given by $P(Y = 1)/PY = 0)$ and take the log to get a number between negative and positive infinity. This is the number we model using our standard regression formula.

(b) Explain what an **odds ratio** means in logistic regression.

(c)) Explain what the coefficients in a logistic regression tell us (i) for a continuous predictor variable and (ii) for an indicator variable.

Solution for (b) and (c): The coefficient β_1, for a variable, X_1, in a logistic regression gives (i) the change in log odds of Y associated with a one-unit change in X_1, assuming all other variables are held fixed, for continuous variables and (ii) the difference in log odds between having and not having a given characteristic for an indicator variable, all else equal. The odds ratio for a variable, X_1 in a logistic regression gives the corresponding impact of X_1 on the odds that Y=1. For instance, suppose that Y is whether or not you get lung cancer, X_1 is your age and X_2 is an indicator for whetehr or not you smoke. Then the odds ratio for X_1 gives the increase in the odds of getting lung cancer associated with being a year older, assuming you have adjusted for smoking status, while the odds ratio for X_2 gives the relative chances of getting lung cancer for smokers versus non-smokers, assuming age has been adjusted for.

(2) Cardiovascular Disease (Based on Rosner 13.58-61): Sudden death is an important, lethal cardiovascular endpoint. Most previous studies of risk factors for sudden death have focused on men. Looking at this issue for women is important as well. For this purpose, data were used from the Framingham Heart Study. Several potential risk factors, such as age, blood pressure and cigarette smoking are of interest and need to be controlled for smilutaneously. Therefore a multiple logistic regression was fitted to these data as shown below. The response is 2-year incidence of sudden death in females without prior coronary heart disease.

Risk Factor	Regression Coefficient (b_j)	Standard Error ($se(b_j)$)
Constant	-15.3	
Blood Pressure (mm Hg)	.0019	.0070
Weight (% of study mean)	-.0060	.0100
Cholesterol (mg/100 mL)	.0056	.0029
Glucose (mg/100 mL)	.0066	.0038
Smoking (cigarettes/day)	.0069	.0199
Hematocrit (%)	.111	.049
Vital capacity (centiliters)	-.0098	.0036
Age (years)	.0686	.0225

(a) Assess the statistical significance of the individual risk factors and explain the practical implications of your findings.

Solution: To get the significance levels for each of these risk factors we need to compute the corresponding Z statistics and either compare them to a critical value or compute the 2-sided p-values. We are given the values of the coefficients and standard errors so this is easy. For instance, for blood pressure we have

$$Z = \frac{b_1 - 0}{s_{b_1}} = \frac{.0019}{.0070} = .27$$

For $\alpha = .05$ the corresponding critical value is our old friend $Z_{.025} = 1.96$. Obviously our test statistic is smaller than the critical value so it does not appear that blood pressure is a significant predictor of sudden death after accounting for the other risk factors. If we wanted the p-value it would be

$$2P(Z \geq .27) = .7871$$

I give the corresponding Z statistics and p-values for the other variables below:

Risk Factor	Z	p-value
Blood Pressure (mm Hg)	.27	.7871
Weight (% of study mean)	-.60	.5485
Cholesterol (mg/100 mL)	1.93	.0536
Glucose (mg/100 mL)	1.74	.0819
Smoking (cigarettes/day)	.35	.7623
Hematocrit (%)	2.27	.0235
Vital capacity (centiliters)	-2.72	.0065
Age (years)	3.05	.0023

It appears that after adjusting for the other factors the only variables that are significant are hematocrit, vital capacity and age, although cholesterol and glucose levels are fairly close to significant. Thus these are the factors that are most important for predicting whether a woman without prior coronary heart disease is at risk for sudden death.

(b) Give brief interpretations of the age and vital capacity coefficients.

Solution: The age coefficient $b_8 = .0686$ means that after holding all the other factors fixed (weight, smoking status, cholesterol levels, etc.) for every extra year of age the log odds of a woman's risk of sudden death goes up by .0686. This is rather hard to interpret since we don't usually think in terms of log odds. It will make more sense in part (c) when we look at odds ratios! For the vital capacity coefficient, $b_7 = -.0098$ we conclude that all else equal, for every extra centiliter of vital capcity, a woman's log odds of sudden death goes **down** by .0098. This is the interpretation of the negative sign. It seems age increases your risk of

sudden death but having extra vital capacity decreases the risk–hardly a surprise!

(c) Compute the odds ratios relating the additional risk of sudden death associated with (i) a 100-centiliter decrease in vital capacity and (ii) an additional year of age after adjusting for the other risk factors.

Solution: The odds ratio for a change Δ in a continuous variable in a logistic regression is given by

$$e^{b_j \Delta}$$

I'll take the age variable first since it is actually simpler. For an increase of 1 year (1 unit) in age, the odds ratio is

$$e^{.0686(1)} = 1.07$$

Thus we conclude that your odds of sudden death get 1.07 times higher for each additional year of age, or increase by 7% per year.

For the the 100 centiliter decrease in vital capcity our change is $\Delta = -100$ so our odds ratio is

$$e^{(-.0098)(-100)} = e^{.98} = 2.66$$

Thus losing 100 centiliters of vital capcity increases your odds of sudden death by a factor of 2.66! This is a big change.

(d) Provide 95% confidence intervals for the odds ratios in part (c)

Solution: Confidence intervals for the odds ratio for a continuous variable are given by

$$[e^{(b_j - Z_{\alpha/2} se(b_j))\Delta}, e^{(b_j + Z_{\alpha/2} se(b_j))\Delta}]$$

For 95% confidence intervals $Z_{\alpha/2} = 1.96$. Plugging in the coefficients and standard errors we get for the age variable

$$[e^{(.0686-(1.96)(.0225))(1)}, e^{(.0686+(1.96)(.0225))(1)}] = [1.025, 1.119]$$

We note that the odds ratio is entirely above 1 meaning that we can be 95% (really 97.5%) sure a woman's odds of sudden death increase with increasing age.
For the vital capacity variable we have

$$[e^{(-.0098-(1.96)(.0036))(-100)}, e^{(-.0098+(1.96)(.0036))(-100)}] = [1.316, 5.396]$$

(e) Predict the probability of sudden death for a 50 year old woman with systolic blood pressure of 120 mmHg, a relative weight of 100% a cholesterol level of 250 mg/100mL, a glucose level of 100 mg/100mL, a hematocrit of 40%, and a vital capacity of 450 centiliters who smokes 10 cigarettes per day. (Note that these numbers are near average for a healthy woman except for the cholesterol level which is high, and of course the number of cigarettes smoked.)

Solution: Plugging in the given values will give us the log odds of sudden death for such a woman:

$$ln(Odds) = -15.3 + .0019(120) - .006(100) + .0056(250) + .0066(100) + .0069(10) + .111(40) - .0098(450) + .0686(50) = -10.083$$

To get the actual probability of sudden dealth we note that

$$P(Y=1) = \frac{e^{ln(ODDS)}}{1+e^{ln(ODDS)}} = \frac{e^{-10.083}}{1+e^{-10.083}} = .00004178$$

This probability is very low which should not be surprising since the risk of sudden death should be small in women with no previous coronary heart disease. None-the-less this number is higher than what we would get if the woman had lower cholesterol or did not smoke.

(3) Ear Infections (Based on Rosner 13.66): In this problem we assess the impact of two different antibiotics on the chances a child will be cured of an ear infection after adjusting for agd and whether one or both ears were infected. The variables are "Clear"–whether the infection has been cleared from both ears after 14 days treatment, "Antibiotic"–the treatment type (1 = Ceftriaxone, 0 = Amoxicillin), Age (categories under two years old, 2-5 years old and 6 year or older), and "NumEars"–the number of ears infected (either 1 or 2). STATA outputs for the pertinent logistic regression model are below. There are two versions, **logit** which gives the raw coefficients and their standard errors and **logistic** which gives the odds ratios and their standard errors.

```
. logit Clear Antibiotic NumEars TwoToFive SixPlus
Logistic regression                               Number of obs   =        203
                                                  LR chi2(4)      =      21.79
                                                  Prob > chi2     =     0.0002
Log likelihood = -129.75295                       Pseudo R2       =     0.0775
------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   .6692876   .3008256     2.22   0.026     .0796802    1.258895
     NumEars |   .0439546    .321911     0.14   0.891    -.5869793    .6748885
   TwoToFive |   1.148698   .3715113     3.09   0.002     .4205494    1.876847
     SixPlus |    1.65964   .4421503     3.75   0.000     .7930418    2.526239
       _cons |  -1.417179   .6001296    -2.36   0.018    -2.593411   -.2409466
------------------------------------------------------------------------------

. logistic Clear Antibiotic NumEars TwoToFive SixPlus
Logistic regression                               Number of obs   =        203
                                                  LR chi2(4)      =      21.79
                                                  Prob > chi2     =     0.0002
Log likelihood = -129.75295                       Pseudo R2       =     0.0775
------------------------------------------------------------------------------
       Clear | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   1.952846    .587466     2.22   0.026     1.082941    3.521528
     NumEars |   1.044935    .336376     0.14   0.891     .5560043    1.963814
   TwoToFive |   3.154084   1.171778     3.09   0.002     1.522798    6.532873
     SixPlus |    5.25742    2.32457     3.75   0.000     2.210109    12.50638
------------------------------------------------------------------------------
```

(a) Overall do these variables help explain how likely a child is to have their ear infections cleared in 14 days? Briefly justify your answer.

Solution: In a logistic regression the likelihood ratio chi-squared test (labeled LR chi2 in STATA) is the equivalent of the overall F test. Here the corresponding p-value is .0002, highly significant, so it seems at least one of antibiotic type, age, and number of ears infected affects how likely a child is to have their ear

infection resolved within 14 days.

(b) Do these variables explain a lot the "variability" in how likely an ear infection is to clear? Explain briefly. What are the practical implications of this statement for treating ear infections in small children with antibiotics?

Solution: In logistic regression the pseudo R-squared plays the role of R-squared and R-squared adjusted in linear regression although you should be cautious in interpreting it as a percentage of variability-it is better jsut to think of it as an index between 0 and 1 assessing model performance with 1 being good. Here the pseudo R-squared is .0775 which is very weak suggesting that the type of antibiotic used is not a very important determinant of the outcome–there must be many other factors we need to know. In fact scientists believe that children with ear infections are treated much too often with antibiotics and that the ear infections would frquently resolve on their own. We can not tell this from our data though as we have no controls who were not given an antibiotic.

(c) Describe what you think would happen if you used backwards stepwise selection to find the best model for predicting whether a child's ear-infection would clear. That is, say what variables would be included in the intial model, what would happen at each step, and what you think the final model would be, and what you would have to do to verify your answer.

Solution: Backwards stepwise model selection works by first fitting the model with all the predictors and the removing the least significant one until all remaining variables are significant. Here we would start with the model given in the problem statement. The only variable that is not significant is number of ears with a p-value of .891. Thus we would remove it from the model first. Most likely after we do that we will be finished because all the other variables are currently highly significant. However, we have to verify this because the current p-values for the other variables assume the presence of number of ears in the model and they could change. The printout for the model with number of ears removed is below for reference through obviously you couldn't do this during an exam. All the remaining variables are indeed significant so this is indeed our final model.

```
. logit Clear Antibiotic TwoToFive SixPlus

Iteration 0:   log likelihood =  -140.6473
Iteration 1:   log likelihood =  -129.8349
Iteration 2:   log likelihood = -129.76231
Iteration 3:   log likelihood = -129.76228

Logistic regression                                     Number of obs   =        203
                                                        LR chi2(3)      =      21.77
                                                        Prob > chi2     =     0.0001
Log likelihood = -129.76228                             Pseudo R2       =     0.0774

------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   .6735338   .2992697     2.25   0.024      .086976    1.260092
   TwoToFive |    1.13779    .362597     3.14   0.002     .4271126    1.848467
     SixPlus |   1.645481   .4296477     3.83   0.000     .8033871    2.487575
       _cons |  -1.350623   .3487325    -3.87   0.000    -2.034126   -.6671198
------------------------------------------------------------------------------
```

(d) Explain briefly how you could figure out what variable to add first in a forwards stepwise model selection procedure for this data.

Solution: In forward stepwise you start with no variables in the model and at each step add the next most useful variable until there is nothing that would be significant when added. In the first stage you have to look at all 1 variable models and see which has the best p-value etc. The way I have set the variables up it seems as if there are four predictors (antibiotic, number of ears, age 2-3 and age 6+). However the indicators for age range are really a group and so you can argue should really be entered together. Note that if you enter one but not the other you are comparing the indicate group to the other TWO groups since one indicator will not differentiate all three groups. Below I have fit the single variable models and the one with both age range variables. It appears that age is the best variable to add first. If you look at the combined age variables the overall p-value is .0002. If you add the indicators sequentially it seems that whether or not you are over 6 is the more important question (corresponding p-value about .01–do note that the overall p-value and the p-value for the individual predictor are NOT exactly the same here which is different from in SLR.) Note that I could also have used contingency table analyses to assess the relationship of any of these variables to whether or not the infection clears!

```
. logit Clear Antibiotic

Logistic regression                               Number of obs   =        203
                                                  LR chi2(1)      =       4.23
                                                  Prob > chi2     =     0.0396
Log likelihood = -138.53076                       Pseudo R2       =     0.0150

------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  Antibiotic |   .5815617   .2841999     2.05   0.041       .02454    1.138583
       _cons |  -.3541718   .2062616    -1.72   0.086    -.7584371    .0500935

. logit Clear NumEars

Logistic regression                               Number of obs   =        203
                                                  LR chi2(1)      =       0.56
                                                  Prob > chi2     =     0.4533
Log likelihood = -140.36611                       Pseudo R2       =     0.0020

------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     NumEars |  -.2184641   .2916499    -0.75   0.454    -.7900873    .3531591
       _cons |   .2497166    .422886     0.59   0.555    -.5791246    1.078558
------------------------------------------------------------------------------

. logit Clear TwoToFive

Logistic regression                               Number of obs   =        203
                                                  LR chi2(1)      =       1.74
                                                  Prob > chi2     =     0.1867
Log likelihood = -139.77538                       Pseudo R2       =     0.0062
```

```
------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   TwoToFive |   .3719711   .2822742     1.32   0.188    -.1812762    .9252184
       _cons |  -.2273898   .1955141    -1.16   0.245    -.6105904    .1558107
------------------------------------------------------------------------------

. logit Clear SixPlus


Logistic regression                               Number of obs   =        203
                                                  LR chi2(1)      =       6.36
                                                  Prob > chi2     =     0.0117
Log likelihood = -137.46978                       Pseudo R2       =     0.0226

------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     SixPlus |   .8471743   .3424626     2.47   0.013     .1759599    1.518389
       _cons |  -.2464004   .1618646    -1.52   0.128    -.5636491    .0708483
------------------------------------------------------------------------------

. logit Clear TwoToFive SixPlus


Logistic regression                               Number of obs   =        203
                                                  LR chi2(2)      =      16.61
                                                  Prob > chi2     =     0.0002
Log likelihood = -132.34405                       Pseudo R2       =     0.0590

------------------------------------------------------------------------------
       Clear |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   TwoToFive |   1.109662   .3574388     3.10   0.002     .4090949    1.810229
     SixPlus |   1.565855   .4211782     3.72   0.000     .7403606    2.391349
       _cons |  -.9650809   .2937848    -3.28   0.001    -1.540889   -.3892732
------------------------------------------------------------------------------
```

(e) Which of the age categories have I used as the reference in this model?

Solution: I have used the "under 2" age category as my reference. Indicators for the other two groups appear in the printout for my logistic regression.

(f) Give brief interpretations of the odds ratios for the "Antibiotic" and "TwoToFive" Variables and show how you would compute them from the information given in the first (logit) printout.

Solution: The odds ratio for the "antibiotic variable" is 1.95 meaning that after controlling for age and number of ears infected, children getting ceftriaxone (the group coded 1) have odds in favor of their infection clearing in 14 days of nearly twice that as children treated with amoxocillin (the reference antibiotic). This interpretation is straightforward because we are dealing with an indicator variable. Similarly, the odds ratio

of 3.15 for the "2 to 5" age group means that after controlling for antibiotic type and number of ears infected, children 2-5 years old have odds in favor of their infections clearing of over 3 rimes as great. This seems like a big improvement. To get these odds ratios we simply take the regression coefficients from the first printout and exponentiate them:

For the antibiotic variable

$$e^{b_1} = e^{.669} = 1.95$$

For the "2 to 5" age variable

$$e^{b_1} = e^{1.14} = 3.15$$

(g) Verify the calculation of the confidence interval for the coefficient of the SixPlus coefficient in the first model and show how to convert it into the confidence interval for the odds ratio given in the second printout.

Solution: In logistic regression we use the normal distribution. For a 95% confidence interval we want $Z_{.025} = 1.96$. Thus the confidence interval for the "6 plus" age group is just $b_4 \pm Z_{\alpha/2} s_{b_4} = 1.66 \pm (1.96)(.442) = [.793, 2.536]$. Since this is an indicator variable, the confidence interval for the odds ratio is obtained by exponentiating the confidence interval for the logistic regression coefficient:

$$[e^{.793}, e^{2.526}] = [2.21, 12.51]$$

(h) According to this model is their a difference in efficacy between Ceftriaxone and Amoxicillin? Write out the details of the appropriate hypothesis test using $\alpha = .05$ (hypotheses mathematically and in words, test statistic, p-value, conclusions.) Does our model show whether either antibiotic helps cure ear infections? Explain briefly.

Solution: We need to test whether the coefficient of the Antibiotic variable is significantly different from 0, namely

$\beta_1 = 0$–When age and number of ears infected have been adjusted for there is no difference in efficacy between the two antibiotics.
$\beta_1 \neq 0$–Even age and number of ears infected have been adjusted for there is a significant difference between the two antibiotics.

Our test statistic is

$$Z_{obs} = \frac{b_1 - 0}{s_{b_1}} = \frac{.669}{.301} = 2.22$$

And the corresponding 2-sided p-value is

$$2P(Z \geq 2.22) = 2(.5 - .4868) = 2(.0132) = .0264$$

Since this p-value is snaller than $\alpha - .05$ we reject the null hypothesis and conclude that there is a difference between the two medications even after adjusting for age and number of ears infected. Specifically, as we saw in part (d) it looks as if ceftriaxone does better than amoxicillin. However technically since we have no control group and we don't even know whether the study was randomized we can not tell whether either medication increases the chance of the ear infection resolving–all we can say is that children on ceftriaxone did better than those on amoxicillin.

(i) According to this model does whether one or both of a child's ears are infected affect their chance of being cured within 14 days using $\alpha = .05$? You do not need to write out the details. Just briefly jsutify your answer.

Solution: The p-value for the "number of ears infected" variable is an enormous .893. Therefore, after adjusting for age and antibiotic, there is no evidence to suggest that it matters how many ears are infected in determining how likely the infection(s) are to resolve within 14 days.

(j) After adjusting for the other factors, does age impact the likelihood of an infection clearing within 14 days? Explain briefly using $\alpha = .05$.

Solution: On the other hand, it appears that age does have an impact in determining how likely the infection is to clear. The p-values for both age group indicators are significant (,002 for "2 to 5" and .000 for "6 plus".) In fact, since the coefficients for these variables are positive and that of "6 plus" is higher than "2 to 5" it seems the older the child is the more likely the infection is to resolve within 14 days, all else equal. However....

(k) Is there a difference in likelihood of cure between children who are 2-5 and children 6 or older? Explain briefly. (Note: I did not refit the model with a different reference group for age–the information you need to get at least an approximate answer is on the printout.)

Solution: If we look at the confidence intervals for the these two age group indicators comparing them to the reference "under 2" group they strongly overlap. This means we can not be sure there is a difference between the "2 to 5" and "6 plus" groups. It seems that what we can be sure of is that being an infant under 2 means your odds of having your infection clear quickly is lower. To be sure about this conclusion we actually would need to refit since the comparison of the itnervals is conservative. However here the overlap is strong enough that the answer is unlikely to change....

(4) Special Delivery: In the developed world most people with HIV receive some form of "highly active antiretroviral therapy" or HAART. (HAART regimens are basically cocktails of multiple drugs that are more effective because the virus is less likely to become resistant in their presence.) However in underdeveloped nations HAART is rarer because of its cost. Professor Helpful believes that HAART regimens will help reduce the risk of HIV positive pregnant women passing on the infection to their babies and must therefore be agressively promoted in poor countries. He has followed n=300 HIV positive pregant women, 100 of whom are receiving at most a basic non-HAART treatment, 100 of whom are taking HAART regimen A, and 100 of whom are taking HAART regimen B. (I'll skip the drug names to keep this simple!) He records Y, whether or not the baby is HIV positive (1 = yes, 0 = no) and which treatment regimen the mother was on ($X_1 = 1$ if the mother was on HAART A and 0 otherwise, $X_2 = 1$ if mother was on HAART B and 0 otherwise), and fits a logistic regression. The corresponding STATA printouts are below. Use them to answer the following questions.

```
. logit HIVplus HAART_A HAART_B


Logistic regression                                     Number of obs   =        300
                                                        LR chi2(2)      =       6.75
                                                        Prob > chi2     =     0.0342
Log likelihood =  -96.32681                             Pseudo R2       =     0.0339

------------------------------------------------------------------------------
     HIVplus |   Coef.      Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |  -0.539        .431      -1.25   0.211    -1.383       0.305
     HAART_B |  -1.286        .534      -2.41   0.016    -2.332      -0.240
```

```
       _cons |  -1.658        .273      -6.08   0.000    -2.193      -1.124
------------------------------------------------------------------------------

. logistic HIVplus HAART_A HAART_B

Logistic regression                               Number of obs   =        300
                                                  LR chi2(2)      =       6.75
                                                  Prob > chi2     =     0.0342
Log likelihood =  -96.32681                       Pseudo R2       =     0.0339

------------------------------------------------------------------------------
     hivplus | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |   .583      .251         -1.25   0.211       .251    1.357
     HAART_B |   .276      .147         -2.41   0.016       .097    0.787
------------------------------------------------------------------------------
```

(a) Overall, is treatment regimen useful for explaining whether a woman passes on HIV infection to her baby? Write down the mathematical hypotheses you are testing, circle the relevant p-value on one of the printouts and give your real-world conclusions using $\alpha = .05$. You do NOT need to provide any other details.

Solution: The equivalent of the overall F test for logistic regression is the likelihood ratio chi-squared test. The hypotheses are just like the F test:

$H_0 : \beta_1 = \beta_2 = 0$-What treatment the mother receives (X_1 and X_2) is not useful for explaining the risk that the mother transmits HIV to her baby. There is no difference in risk among any of the groups.
H_A : At least one of β_1, β_2 is non 0; at least one of the treatment group variables is useful for explaining risk. There is a difference in risk among the non-HAART, HAART A and HAART B regimens.

The test statistic is $\chi^2 = 6.75$ and the corresponding p-value is .0342. Since this is less than $\alpha = .05$ we reject the null hypothesis and conclude that risk of transmission does differ by treatment group. This model does overall help to explain how likely a mother is to transmit HIV to her baby. Note that per the problem statement all you needed were the math hypotheses, the p-value and the conclusions. You did not need to write out the word hypotheses or test statistic but I included them as a study aid since I certainly could ask for them!

(b) Give a brief interpretation of the odds ratio for the HAART A variable and show how to compute it from the first regression printout.

Solution: In a logistic regression, the Odds Ratio for an indicator variable tells you how much higher (or lower) the odds of an event (Y=1, here mother transmits HIV to infant) are for someone who has the characteristic of interest (HAART A treatment) than someone who doesn't (here the reference group, no HAART), all else equal. An odds ratio of 1 corresponds to equal odds, an odds ratio above 1 means the odds are higher for the person with the characteristic, and an odds ratio below 1 means the odds are lower for the person with the characteristic. From the second printout the odds ratio for the HAART A variable is OR = .583, meaning the odds the mother transmits HIV to her baby are only 58.3% or a little over half as high for a mother receiving HAART A treatment as for a mother not receiving any HAART treatment. (Note that here we don't need the all else equal because a mother can be in only 1 group and there are no other predictors.) To compute the odds ratio we simply exponentiate the corresponding regression coefficient:

$$OR_A = e^{b_1} = e^{-.539} = .583$$

(c) Do HAART A and HAART B appear to **reduce** a mother's risk of passing on HIV to her infant? Explain briefly using $\alpha = .05$ and give the p-values corresponding to the tests you are performing. You do NOT need to write out any other details of the tests.

Solution: Our best estimate is that both HAART A and HAART B reduce the risk of mother to infant transmission relative to the non-HAART group since the odds ratios for both the group indicators are below 1. However to be sure we need to look at the p-values for the 1-sided tests that $\beta < 0$. We get these p-values by dividing the STATA p-values for the corresponding Z tests in half. The p-value for the HAART A indicator is ..105, well above .05, meaning we can not be sure that the coefficient for the HAART A indicator is negative or correspondingly that the odds ratio is significantly below 1. Therefore, we can not be 95% sure the treatment is associated with reduced risk. The p-value for HAART B on the other hand is $.016/2 = .008$, which is below .05 meaning we can be sure the risk in the HAART B group is significantly lower than in the no-HAART group. Here I specifically asked you to give the p-values. However in general you could also answer this sort of question using confidence intervals. There are two ways to do this. First, we could check whether the confidence intervals for the odds ratios are entirely below 1. If they are then the risk is lower on the treatments than on the control regimen. Second, we could use the confidence intervals for the coefficients (on the logit scale) and check whether or not they are entirely below 0. Do keep in mind however that the confidence intervals are two-sided so it is actually a tougher standard to say the whole 95% CI has to be below a given value than it is to do a 1-sided test that you are below a given value.

(d) Find the odds ratio comparing the risk of HIV transmission for mothers in the HAART A group compared to those in the HAART B group. Show your work. Based on this estimate which of these treatment regimens is more effective? Briefly explain your reasoning. Do you think you can be 95% sure this treatment is better? Explain.

Solution: There are several ways to do this. The easiest way is to note that the odds ratio for HAART A versus no HAART is $ODDS_A/ODDS_{None}$ and the odds ratio for HAART B versus no HAART is $ODDS_B/ODDS_{None}$ so if we take the ratio of these odds ratios we get the odds ratio for A vs B:

$$\frac{OR_{AvsNone}}{OR_{BvsNone}} = \frac{ODDS_A/ODDS_{None}}{ODDS_B/ODDS_{None}} = \frac{ODDS_A}{ODDS_B} = OR_{AvsB}$$

For these data we get $OR_{AvsB} = .583/.276 = 2.11$ Since this number is greater than 1, it appears the risk of transmission is higher in the HAART A group than the HAART B group. Thus the HAART B treatment appears to work better. However, the confidence intervals for the odds ratios comparing each of HAART A and HAART B to no HAART overlap with each other. Thus it is possible that the two treatments are associated with the same improvement in risk relative to no HAART and we can NOT be 95% sure that HAART B is better.

(5) Prenatal Care-acteristics: Professor Helpful recognizes that there are probably many factors besides treatment regimen that affect whether a mother transmits HIV to her baby. He has thus added the following variables to his logistic regression model from Question 4: X_3, the mother's viral load in copies per milliliter of blood (higher viral load is worse), X_4, the mother's age in years, X_5, the number of years the mother has been HIV positive, X_6, the number of weeks during the pregnancy for which the mother was receiving HAART therapy, and X_7 the method by which the baby was delivered (1 = C-section, 0 = natural delivery). The new printouts are given below. Use them to answer the following questions.

```
. logit HIVplus HAART_A HAART_B VLoad Age YrsHIV WksHAART Delivery

Logistic regression                               Number of obs   =       300
```

```
                                                LR chi2(7)      =     32.47
                                                Prob > chi2     =     0.000
Log likelihood =  -26.51722                     Pseudo R2       =     0.500
------------------------------------------------------------------------------
     HIVplus |   Coef.      Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |  -0.70        0.250       2.80  0.005      [-1.19,   -0.21]
     HAART_B |  -1.80        0.300       6.00  0.000      [-2.39,   -1.21]
       VLoad |0.00001    0.0000025       4.00  0.000    [.000005, .000015]
         Age |   0.10        0.050       2.00  0.046      [ 0.00,    0.20]
      YrsHIV |   0.10        0.080       1.25  0.211      [-0.06,    0.26]
    WksHAART |  -0.05        0.010      -5.00  0.000      [-0.07,   -0.03]
    Delivery |  -0.40        0.150      -2.67  0.004      [-0.69,   -0.11]
       _cons |  -5.00        0.500     -10.00  0.000      [-5.98,   -4.02]
------------------------------------------------------------------------------

. logistic HIVplus HAART_A HAART_B

Logistic regression                             Number of obs   =       300
                                                LR chi2(7)      =     32.47
                                                Prob > chi2     =     0.000
Log likelihood =  -26.51722                     Pseudo R2       =     0.500
------------------------------------------------------------------------------
     HIVplus |    OddsRatio       z      P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
     HAART_A |      0.4966       2.80    0.005        [0.3042,  0.8106]
     HAART_B |      0.1652       6.00    0.000        [0.0916,  0.2982]
       VLoad |     1.00001       4.00    0.000     [1.000005, 1.000015]
         Age |      1.1052       2.00    0.046        [1.0020,  1.2190]
      YrsHIV |      1.1052       1.25    0.211        [0.9448,  1.2928]
    WksHAART |      0.9512      -5.00    0.000        [0.9328,  0.9700]
    Delivery |      0.6703      -2.67    0.004        [0.4996,  0.8994]
------------------------------------------------------------------------------
```

(a) Find the probability that a 30 year old women on HAART A for 20 weeks of her pregnancy with a viral load of 10,000 who has been HIV positive for 10 years will have an HIV **negative** baby if she delivers by Cesarean Section. Show your work.

Solution: First note that the model predicts the probability of having an HIV **positive** baby so once we get the predicted probabiltiy we will have to subtract it from 1 to get our final answer! The formula for the predicted probability is

$$p = \frac{e^{b_0+b_1X_1+\cdots+b_kX_k}}{1+e^{b_0+b_1X_1+\cdots+b_kX_k}}$$

It is easiest to first plug in the X values and then do the exponentiation. Since the mother is on HAART A, $X_1 = 1$ and $X_2 = 0$. Since she delivers by C-section, $X_7 = 1$. All the other numbers are straightforward. We have

$$b_0 + b_1X_1 + \cdots + b_kX_k = -5 - .7(1) - 1.8(0) + .00001(10000) + .1(30) + .1(10) - .05(20) - .4(1) = -3$$

From this we get the probability as

$$p = \frac{e^{-3}}{1 + e^{-3}} = .047$$

Thus the probability that a woman with the given characteristics will have an HIV positive baby is 4.7% and the chance she will have an HIV negative baby is 95.3%.

(b) Explain as precisely as you can the meaning of the p-value for X_7, the delivery variable. Your answer should be specific to this context and incorporate the relevant numeric value(s).

Solution: The p-value is, in general, the probability of getting data as or more extreme (i.e. as or more favorable to H_A) than what was observed, assuming the null hypothesis is true. Here that translates to saying that there is only a 4 out of 1000 chance (numerical value of p-value= .004) that we would have seen such a big difference in HIV transmission rates in our sample between woman who had C-sections version woman who did not (our data) if really the delivery method was not associated with risk of transmission after adjusting for the other factors.

(c) (i) Give a brief interpretation of the confidence interval for the odds ratio for X_6, the weeks treated variable. (ii) Find a 95% confidence interval for the odds ratio associated with an extra MONTH (4 weeks) of HAART treatment. Based on this latter interval can you be sure that, all else equal, an extra month of HAART treatment will reduce the risk of mother to child transmission by 10%.

Solution: (i) The CI for the odds ratio of the weeks treated variable is [0.9328, 0.9700] which says that your odds of having an HIV positive baby with x+1 weeks of HAART treatment are between 93.28%-97% as high as they would be with only x weeks of HAART treatment. A more natural way to say this is that each additional week of HAART treatment is associated with a decrease of between 3% to 6.72% in the odds of HIV transmission, all else equal. Since this interval is entirely below 1, we are 95% sure that additional time on HAART is associated with a LOWER risk of transmission.

(ii) There are several ways to approach this problem. We are talking now about a 4 unit change in X_6 so we can either multiply the confidence interval for β_6 by 4 to get the change in log odds associated with an extra month of HAART and then exponentiate to get the corresponding CI for the odds ratio, or we can take the current interval for the odds ratio and raise the ends to the 4th power since multiplication on the log odds scale is exponentiation on the odds ratio schedule. Using the first approach, the CI for the change in log odds corresponding to 4 extra weeks of HAART is $[-0.07, -0.03] * 4 = [-0.28, -0.12]$. Exponentiating this gives us $[e^{-.28}, e^{-.12}] = [.76, .89]$. We get the same thing (up to rounding) from raising the current interval for the odds ratio to the 4th power: $[(.9328)^4, (.97)^4] = [.76, .89]$. This says that each additional month of HAART treatment is associated with between an 11% to 24% reduction in the odds. Or if you prefer the odds for a woman with an extra month of HAART are only .76 to .89 as high. Even in the worst case there is at least an 11% reduction in the odds so it looks like we can make the desired claim that an extra month of HAART decreases the odds of transmission by at least 10%.

(d) Professor Helpful believes overfitting is an issue in this model. (i) Explain why he is correct. (ii) Give a possible real-world cause of the overfitting and say how you would check whether your idea was correct. (iii) Say what variable you would remove first in a backwards stepwise procedure and why. (iv) What do you think would happen to the pseudo R^2 if you removed this variable? Why?

Solution: (i) Overfitting means including variables in your model that are not useful. Here the years of HIV variable has a p-value of .211 meaning it is not statistically significant and not worth including the model once we have taken all the other factors into account. (ii) It seems likely that how long the person has had HIV is strongly correlated with their age and their viral load (which measures how sick they are). Thus there is probably an issue of multicollinearity. We could check this by looking at the correlations among the three

variables. (iii) Since years of HIV is the only variable with a non-significant p-value it would be the first thing to be removed in a backwards stepwise procedure. (iv) This actually depends on whether or not the pseudo R-squared used by STATA is an adjusted R-squared that takes degrees of freedom into account. If it is, then my reducing the overfitting the R-squared value might actually go up a little. If it is an unadjsuted value then it would get slightly smaller or stay the same. It can't change very much since the years of HIV variable is not significant and therefore not explaining much of the variability in the outcome. Therefore taking it out can cause very little reduction in our estimate of how much we have explained.

(6) Sports Fanatics: My husband, Gareth, is from New Zealand where the national sports passion is rugby (sort of like American football only better!) The national rugby team is called the All Blacks (they wear black) and their main rivals are Australia (the Wallabies) and South Africa (the Springboks). Gareth realizes that what he really cares about is whether the All Blackswin or not. Therefore he decides to perform a logistic regression with the the response variable, Y, being whether or not the All Blacks win (Y = 1 if they win and 0 if they lose). The predictors are

AB Win%=the percent of the previous ten games that the All Blacks had won going into the game in question, ranging from 0 to 100

OppWin%, (same definition for the opponents last 10 games)

Home?, an indicator variable with 1 corresponding to an All Blacks home game and 0 an away game

Temperature (the temperature at which the game was played.)

Australia? (a dummy variable with 1 corresponding to a game against archrival Australia and 0 a game against another team.)

Below are the p-value for the likelihood ratio chi-square test along with a table of coefficients, standard errors, Z scores and p-values for the various variables. Use them to answer the questions below.

```
LR chi2 p-Value < 0.0001
```

	Coef	SE	Z	p-value
Constant	-25.30	10.54	-2.40	0.0163
AB Win %	0.466	0.176	2.65	0.0082
Opp Win %	-0.170	0.643	-2.65	0.0081
Home?	1.45	0.660	2.20	0.0278
Temperature	0.115	0.045	2.55	0.0108
Australia?	-0.245	1.890	-0.13	0.8969

(a) Is there evidence that at least one of the variables is a statistically significant predictor of whether the All Blacks win? Justify your answer.

Solution: Yes, the p-value for the likelihood ratio chi-squared test is very small (¡0.0001). This indicates that we can reject the null hypothesis that none of the variables are helping to predict wins. At least one variable is a significant predictor. Mathematically our hypotheses would have been $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ versus H_A : At least one $\beta \neq 0$.

(b) What does the coefficient for Temperature tell us about the relationship between Temperature and the probability that the All Blacks win? Compute the corresponding odds ratio for a 10 degree increase in

temperature and explain what it means. Give a confidence interval for this odds ratio.

Solution: The coefficient for temperature is positive. Hence, holding all other variables fixed, on warmer days the All Blacks are more likely to win than on colder days. More specifically, the log odds of an All Blacks victory goes up .115 per degree increase in temperature. These units are hard to understand so we can convert the value to an odds ratio by exponentiating. I didn't ask for this but will include it anyway. We get

$$\hat{O}R_{temp} = e^{.115} = 1.12$$

This means that all else equal the odds of the All Blacks winning go up by a factor of 1.12 for each degree of temperature or 12% per degree. Since this value is above 1, higher temperatures favor the All Blacks. Note that we could also get a confidence interval for this odds ratio by getting a confidence interval for the coefficient and then exponentiating it.

The second part of the question asked for an odds ratio for a temperature jump of 10 degrees. Temperature is a continuous variable so we are talking about a delta of 10 degrees or $\Delta = 10$. The odds ratio for a numeric variable corresponding to a change delta is

$$\hat{O}R_{\Delta} = e^{b*\Delta} = e^{.115*10} = e^{1.15} = 3.16$$

Thus the odds of winning are 3.16 times higher if the temperature is 10 degrees hotter. You can verify that this is the same answer we get by raising the odds ratio for the one degree range to the power 10. If we wanted a confidence interval for this odds ratio we use the formula

$$[e^{(b-Z_{\alpha/2}s_b)\Delta}, e^{(b+Z_{\alpha/2}s_b)\Delta}]$$

Here we have b = .115, Z = 1.96 for a 95% confidence interval, $s_b = .045$ and $\Delta = 10$. Plugging these numbers in gives a CI of [1.307, 7.629]. Thus the odds of winning are somewhere between 1.307 and 7.629 times as high if the temperature goes up 10 degrees, all else being equal.

(c) Which variables are statistically significant? Justify your answer. Do the signs of the various coefficients make sense?

Solution: AB Win% (p-value 0.0082), Opp Win% (p-value 0.0081), Home? (p-value 0.0278) and Temperature (p-value 0.0108) are all statistically significant variables because they have low p-values. The Australia indicator is not significant because it's p-value is above $\alpha = .05$. Note that we could have found these p-values using the Z table if they had not been given to us! As far as the signs, the better the All Blacks have been playing the more likely they are to keep winning so the positive sign on AB Win% makes sense. However if the All Black's oppoent has been playing well it will be a harder game so the chances of winning will go down. Thus the negative sign on Opp Win% makes sense. Similarly home field is an advantage so we would expect the Home? coefficient to be positive as it is. New Zealand is a warm country so it is not surprising the All Blacks play better in warmer weather as suggested by the positive sign on the temperature variable. The All Black's archrival Australia is the 2nd best team in the world (compared to the All Black's of course!) so games agains them are harder and we would expect a negative coefficient. The fact that this isn't significant indicates just how good the All Blacks are! Of course I wouldn't expect you to know these extra rugby facts for our final!

(d) Estimate the probability of the New Zealand All Blacks winning a game against South Africa played in South Africa at 50 degree temperatures where both teams have a winning percentage of 70.

Solution: The formula is

$$p = \frac{e^{b_0+b_1X_1+...+b_5X_5}}{1+e^{b_0+b_1X_1+...+b_5X_5}} = \frac{e^{-25.3+.466(70)-.170(70)+1.45(0)+.115(50)-.245(0)}}{1+e^{-25.3+.466(70)-.170(70)+1.45(0)+.115(50)-.245(0)}} = .763$$

The All Black's chances of winning the game are quite good!

(e) Find a confidence interval for the coefficient of the Home? variable and give a brief interpretation. Also find the odds ratio for the corresponding variable and a 95% confidence interval and interpret those results.

Solution: The CI is just $b_3 \pm Z\alpha/2s_{b_3} = 1.45 \pm (1.96)(.66) = [.1564, 2.7436]$. The log odds of winning is between ,1564 and 2.7436 higher when the game is at home than when it is away, all else equal. Since the whole CI is above 0 we are 95% sure that the All Blacks are more likely to win a home game than an away game, all else equal. To convert this to a CI for the odds ratio we expoentiate. The OR CI is [1.17, 15.54]. This means our odds of winning a home game are 1.17 to 15.54 times as high as for an away game all else equal. Since this whole CI is above 1 again we conclude that a home game is an advantage all else equal.

(f) The coefficient for the Home? variable seems to indicate that the All Blacks are more likely to win at home than on the road. However, somewhat surprisingly, the All Blacks turn out to win more games on the road than at home. One of my husbands MBA students (from that school on the wrong side of town) looks at these results and states that this indicates that there must be some mistake in the analysis. However, you tell them that in fact this apparent inconsistency is entirely possible even if the model is correct. Assuming that the model is correct (i.e. there are no important variables missing from the model or violations of the basic assumptions etc.) and the coefficient estimates are exactly correct how could the coefficient for Home? be positive even though the All Blacks win more games on the road?

Solution: There are at least a couple of possible explanations for this effect. The key here is that the Home? coefficient being positive only tells us that if all the other variables are the same then the All Blacks are more likely to win at home than on the road. However, the other variables may not all be the same. For example, suppose that the All Blacks always play better teams (as measured by Opp Win %) at home and worse teams on the road. Then this might negate the otherwise positive effect of being at home. Another possibility is temperature. Suppose that the All Blacks home games tend to be colder than their away games. Then, since they prefer playing in warmer temperatures, they could well end up winning more games on the road. This is the sort of reason why it can be critical to adjust for possible confounders in a regression model!

Regression/ANOVA Problems

For regression some of the best practice problems are the warm-up problems from homework 5 and 6. You can also take the regression printouts from the midterm 1 practice set and use them to think about the topics we have covered more recently. I have included below an extension about one of those problems which I had previously included only in ANOVA form and added some regression parts. This is the problem as it originally appeared in an exam. I've also taken a problem from the Midterm 1 practice that had a good example of an outlier and added some additional parts to it. Let me know if after going through all those options there is anything on which you feel short of practice....

(1) Analysis of Varying Medications: A researcer is analyzing methods of reducing cholesterol levels. She is interested in the relative merits of diets versus cholesterol lowering medications. For each of 65 subjects who began the study with high cholesterol she records total blood cholesterol level (in mg per deciliter) after 6 months participation in the study. The patients are divided into G=5 groups: a control group (C) which receives a placebo, a vegetarian diet group (V), a low fat diet group (LF), a low dose medication group (LD) and a high dose medication group (HD). STATA printouts below show the group means, standard deviations, and group sizes, along with an ANOVA table which seems to be missing a few numbers. Use this

information to answer the questions on the following pages.

```
. oneway Cholesterol Group, tabulate

           |       Summary of Cholesterol
           |
     Group |        Mean   Std. Dev.       Freq.
-----------+------------------------------------
         C |       240        1.22           25
         V |       225        1.18           10
        LF |       230        1.10           10
        LD |       215        1.02           10
        HD |       200        1.11           10
-----------+------------------------------------
     Total |       226.2     13.51           65

                       Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      _6881.4_       _4_    _1720.4     _21.5_      0.000
 Within groups        _4800_       60      80.00
------------------------------------------------------------------------
    Total            11681.4      _64_
```

(a) Fill in the missing entries (marked ____) in the above tables. You do not need to give your reasoning though it may help if you make mistakes. There is an easy way and a hard way to do this! Try to do it the easy way! Note that you will be able to do the rest of the problem even if you can not do part (a).

Solution: The filled in table is above. The degrees of freedom between is the number of groups minus 1, here G-1 = 5-1 =4. The degrees of freedom total are n-1 = 64 in this case. To get the rest, just use the fact that mean squares are sums of squares divided by degrees of freedom so SSW = MSW*df = 80*60 = 4800. Then note that the sums of squares must add so SSB = SST - SSW = 6881.4. This gives MSB = 6881.4/4 = 1720.4. Finally F = MSB/MSW =21.5.

(b) Based on this data is there evidence that any of the group means are different from each other? Justify your answer by performing an appropriate hypothesis test. Be sure to state the null and alternative hypotheses, both mathematically and in words, give the p-value, and your real-world conclusions.

Solution: We are asked to do an overall F test. In an ANOVA our hypotheses are:

$H_0 : \mu_C = \mu_V = \mu_{LF} = \mu_{LD} = \mu_{HD}$–all the groups have the same average cholesterol level–the diets and medications have no impact.
H_A : The μ's are not all equal–at least two of the means are different from each other–which here would imply at least one of the diets or drugs has an effect (or at least differes from one of the other treatments).

The p-value corresponding to the F statistic is 0 so at $\alpha = .05$ we reject the null hypothesis and conclude that at least one of the groups as a mean that is not the same as the others.

(c) Suppose that instead of doing an ANOVA we had fit a regresison model to this data using the vegetarian diet group as the reference group. Write down the estimated regression equation we would have obtained.

Solution: In an ANOVA setting the intercept represents the mean of the reference group and the other coefficients represent the difference between a particular group and the reference group. Here we are asked to use the vegatarian diet as the reference, so $b_0 = 225$. The difference between the control and vegetarian groups is 240-225 = 15 so we would have $b_c = 15$. Similarly, $b_{LF} = 230 - 225 = 5, b_{LD} = 215 - 225 = -10, b_{HD} = 200 - 225 = -25$. Note that the minus signs for the latter two groups mean that those groups have LOWER cholesterol levels than the people on the vegetarian diet. Our equation is thus

$$\hat{Y} = 225 + 15X_C + 5X_{LF} - 10X_{LD} - 25X_{HD}$$

(d) What percentage of the variability in cholesterol levels is explained by the treatment group to which a subject belonged? Show your calculations or explain your reasoning.

Solution: We are being asked for R^2 which is just SSB/SST in an ANOVA. Here we get SSB/SST = 6881.4/11681.4 = .589 or a little under 60% of the variability is explained by treatment. This is pretty high considering how many things can affect a person's cholesterol level!

Below is a table showing test statistics and p-values for pairwise comparisons of the different group means for this ANOVA. Use it to help answer parts (e)-(f).

Pair	t_{obs}	p-value	Pair	t_{obs}	p-value
C vs V	4.48	.00003	C vs LF	2.99	.00404
C vs LD	7.47	.00000	C vs HD	11.95	.00000
V vs LF			V vs LD	2.50	.01517
V vs HD	6.25	.00000	LF vs LD	3.75	.00040
LF vs HD	7.50	.00000	LD vs HD	3.75	.00040

(e) The test comparing the vegetarian diet group to the low fat diet group is missing. State the null and alternative hypotheses mathematically and in words, compute the test statistic and an approximate p-value and explain your real-world conclusions. (Note: Make sure you carefully show your calculation of the standard error.)

Solution: Our hypotheses are

$H_0 : \mu_V = \mu_{LF}$ or $\mu_V - \mu_{LF} = 0$–the mean cholesterol level of people on the two diets is the same
$H_Z : \mu_V \neq \mu_{LF}$ or $\mu_V - \mu_{LF} \neq 0$–the cholesterol levels of people in the two diet groups are different.

The standard error of the difference in means is

$$se = \sqrt{MSW(\frac{1}{n_V} + \frac{1}{n_{LF}})} = \sqrt{80(\frac{1}{10} + \frac{1}{10})} = \sqrt{16} = 4$$

Our test statistic is therefore

$$t_{obs} = \frac{\bar{Y}_V - \bar{Y}_{LF} - 0}{\sqrt{MSW(\frac{1}{n_V} + \frac{1}{n_{LF}})}} = \frac{225 - 230}{4} = -1.25$$

Under the null hypothesis this test statistic has a t distrobution with n-G = 60 degrees of freedom. Looking in the row on the t-table for 60 degrees of freedom we see that $t_{.85} = 1.046$ and $t_{.90} = 1.296$. Thus the one-sided p-value would be between .1 and .15 and the corresponding 2-sided p-value would be between .2 and .3. This is a very large p-value so we fail to reject the null hypothesis. We do not have sufficient evidence to show a difference in cholesterol levels between the two diet groups.

(f) Which pairs of means are significantly different from one another at the $\alpha = .05$ level without adjusting for multiple testing? Explain briefly.

Solution: Looking at the table, all the p-values are less than .05 except for the one we just calculated for the vegetarian versus low fat diet. Thus all the groups are significantly different except the two diets if we don't adjust for multiple testing.

(g) According to the Bonferroni method, what significance level should you use for the individual tests for differences of means to get an overall significance level of $\alpha = .05$? Explain briefly. Use your answer to repeat part (f), adjusting for multiple comparisons. Indicate any results that have changed.

Solution: The Bonferroni method says that significance level to use for individual tests is the overall desired significance level α divided by the number of tests. Here we have 5 choose 2 = 10 tests (either use the binomial coefficient or just count the number of pairs in the table) and our overall significance level is meant to be $\alpha = .05$ so for the individual tests we should use $\alpha^* = .005$. Looking at the table we see that the test for the vegetarian diet versus the low dose medication has a p-value greater than .005 so we now cannot conclude these two treatments differ. Otherwise all the results remain the same. In summary we can not be sure the vegetarian diet differs from either the low dose medication or the low fat diet but all the other means are significantly different from one another even after adjusting for multiple testing.

(h) The researcher is interested in comparing the average cholesterol level of people in the two diet groups with that of the people in the low dose medication group. Write down an appropriate linear combination, L, for the comparison she wishes to do. Give your best estimate of L and the corresponding standard error and use these numbers to find a 95% confidence interval for L. Give a brief interpretation of your interval and explain whether the researcher can conclude there is a difference in efficacy between diets and the low dose medication in reducing cholesterol levels.

Solution: This is a linear combination problem. We want to compare μ_{LD} with the average of the two diet groups. Thus the combination of interest is

$$L = \mu_{LD} - \frac{\mu_V + \mu_{LF}}{2} = \mu_{LD} - .5\mu_V - .5\mu_{LF}$$

Our best estimate of a linear combination of means is simply to plug in the sample means, $\bar{Y}$. Thus our estimate is

$$\hat{L} = \bar{Y}_{LD} - .5\bar{Y}_V - .5\bar{Y}_{LF} = 215 - .5(225) - .5(230) = -12.5$$

The standard error for a linear combination is

$$se(\hat{L}) = \sqrt{MSW \sum \frac{c_j^2}{n_j}} = \sqrt{80(\frac{1^2}{10} + \frac{(.5)^2}{10} + \frac{(.5)^2}{10})} = 3.46$$

We want a 95% confidence interval and we are working with n-G = 60 degrees of freedom. From the t-table we see that $t_{60,.025} = 2$ Thus our confidence interval is $-12.5 \pm (2)(3.46) = [-19.42, -5.58]$. Thus it seems that the average cholesterol level in the low dose medication group are between 5.58 and 19.42 mg/dL LOWER than the average in the two diet groups. Since this interval lies entirely below 0 we can be 95% (really 97.5%) sure that the patients in the loe dose medication group are doing better than those who are only using dietary measures to control their cholesterol.

Obviously there are factors other than treatment group which could affect a person's cholesterol level. Thus, the researcher has fit a multiple regression of cholesterol level on treatment group, age, weight and whether or not the person has a family history of coronary artery disease (1 = yes and 0=no). Use the STATA

multiple regression printout to answer the remaining parts of the question.

```
. reg Birthweight EX LOW HIGH

      Source |       SS       df       MS              Number of obs =      65
-------------+------------------------------           F( 7 , 57  )  =    73.3
       Model | 10513.3         7   1501.9              Prob > F      =   0.000
    Residual |  1168.1        57     20.5              R-squared     =   0.900
-------------+------------------------------           Adj R-squared =   0.888
       Total | 11681.4        64                       Root MSE      =   4.528

------------------------------------------------------------------------------
 Birthweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         AGE |       0.5       0.2        2.50   0.0153      0.10       0.90
      WEIGHT |       0.6       0.2        3.00   0.0040      0.20       1.00
     HISTORY |      30.0       5.0        5.00   0.0000     20.00      40.00
         VEG |      -2.0       1.6       -1.25   0.2164     -5.20       1.20
     LOW FAT |       1.0       0.8        1.20   0.2351     -0.60       2.60
    LOW DOSE |      -5.0       3.0       -1.67   0.1004    -11.00       1.00
   HIGH DOSE |     -25.0       5.0       -5.00   0.0000    -35.00     -15.00
       _cons |     100.0      25.0        4.00   0.0002     50.00     150.00
------------------------------------------------------------------------------
```

(i) In terms of percentage of variability explained and accuracy of predictions does this model do a better job than the simple ANOVA from parts (a)-(h)? Explain briefly what numbers from the printout you are looking at to answer this question and also perform an appropriate hypothesis test. Does this model make good predictions? Explain.

Solution: In part (d) we found that $R^2 = .589$ for the ANOVA while here we have $R^2 = .9$ and $R^2_{adj} = .888$. Clearly adding the additional variables has improved the percentage of variability exlained. For predictions we must look at the RMSE. For the regression model our average error is RMSE = 4.528. For the ANOVA RMSE $= \sqrt{MSW} = \sqrt{80} = 8.94$. Thus the predictions from the regression are substantially more accurate. To tell if the predictions from the regression model are actually good we compare the errors to the Y values we are trying to predict. We know that normal cholesterol levels are around 200. The grand mean for this data set (from the table at the start of the problem) is 226.2. Based on either of these numbers we are making between a little over a 2% error (e.g. $4.58/200 = 2.29\%$) which seems quite good.

We were also asked to do a formal test to determine whether the new model is better than the old one. What is needed is a partial F test. We have added three variables, age, weight and family history. Thus our hypotheses are:

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$–none of age, weight and family history help explain cholestoral level.
H_A : At least one of those β's is not 0. The model with the extra variables is a significant improvement. Our test statistic is

$$F = \frac{(SSR_{full} - SSR_{red})/3}{SSE_{full}/(n-7-1)} = \frac{(10513.3 - 6881.4)/3}{1168.1/57} = 59.08$$

Technically we need an F table or a STATA printout of the partial F test to get the p-value but by now your intuition should be telling you that this is a huge value and that we will reject the null hypothesis. The new

model is defintiely an improvement over the one involving only the treatment groups.

(j) After adjusting for age, weight, and family history, does it appear that the diets or medication doses have a significant impact on cholesterol levels compared to the control group? Briefly justify your answer.

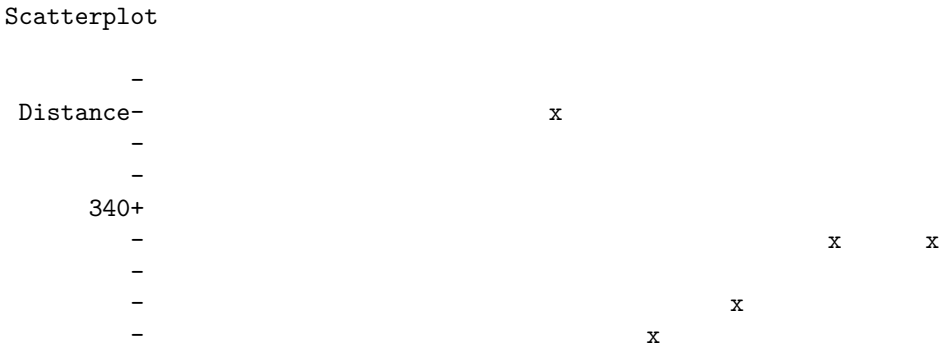
Solution: The indicators for the V, LF, LD, and HD groups compare the cholesterol levels for those groups to the reference group (here the controls) after adjusting for the other variables. Here the only one of those variables that is significant at $\alpha = .05$ is the high dose medication indicator. All the others have p-values over .1. Thus the only treatment with a significant impact appears to be the high dose medication.

(k) Your answer to part (j) is different from what you found in parts (f) and (g). Explain what has happened and what it implies about whether the researcher performed a properly randomized study.

Solution: Before ALL the treatments looked like they worked. After adjusting for age, weight and family history, only one seems to work. This would suggest that there were differences in age, weight or family history between the different treatment groups–otherwise the adjustment shouldn't have changed anything. This is bad because it means that the randomization was not properly done. The whole point of randomization is to balance out possible confounding factors so that they do not affect the conclusions about the treatments. For instance, here, maybe more overweight people ended up in the control group, making it look worse than it really was. Or more young people ended up in the diet groups making them look better than they really were, etc.

(2) Leaping Into the Future: In the modern Olympic era, performances in track and field have been steadily improving. The table below gives the winning distance (in inches) for the Olympic long jump from 1952 to 1984. Below is a regression printout for a simple regression of distance on year. Use the printout to answer the following questions.

Year	Distance
1952	298
1956	308.25
1960	319.75
1964	317.75
1968	350.5
1972	324.5
1976	328.5
1980	336.25
1984	336.25



Regression Analysis

```
. reg Distance Year

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =    9.21
       Model |  1137.52604     1  1137.52604           Prob > F      =  0.0190
    Residual |  864.973958     7  123.567708           R-squared     =  0.5681
-------------+------------------------------           Adj R-squared =  0.5063
       Total |      2002.5     8    250.3125           Root MSE      =  11.116

------------------------------------------------------------------------------
    Distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Year |   1.088542   .3587706     3.03   0.019     .2401839    1.936899
       _cons |  -1817.833   706.0703    -2.57   0.037    -3487.424   -148.2423
------------------------------------------------------------------------------

. reg Distance Year Yearsq

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  2,     6) =    6.88
       Model |   1394.3493     2  697.174648           Prob > F      =  0.0280
    Residual |  608.150703     6  101.358451           R-squared     =  0.6963
-------------+------------------------------           Adj R-squared =  0.5951
       Total |      2002.5     8    250.3125           Root MSE      =  10.068

------------------------------------------------------------------------------
    Distance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Year |   225.7233   141.1208     1.60   0.161    -119.5868    571.0333
      Yearsq |  -.0570718   .0358538    -1.59   0.163    -.1448028    .0306591
       _cons |  -222852.3   138860.1    -1.60   0.160    -562630.8    116926.1
------------------------------------------------------------------------------
```

****

(a) Give the units and interpretation of b_1 in the simple regression model.

Solution: The regression coefficient b_1 always gives than change in Y associated with a one unit change in X. Since b_1 must convert from X units to Y units, the units of b_1 are the units of Y divided by the units of X. In this problem, X is in years and Y is distance in inches, so the units of b_1 are inches per year. Since $b_1 = 1.08854$, a one unit change in year is associated with a 1.08854 inch change in distance, i.e. the winning long jump distance increases by 1.08854 inches per year. Naturally, since the Olympics are only held every four years, this really means that the winning distance increases by about 4.35 inches every Olympiad.

(b) What proportion of the variability in distance is explained by year using the simple linear regression model? Does the model do a good job in this respect?

Solution: The proportion or percentage of variability explained by the regression is given by $R^2 = 56.8\%$, or, if we want an unbiased estimate, by $R^2_{adj} = 50.6$. Whichever number you use, the regression is explaining barely over half the variability and leaving nearly half the variability unexplained. This is not very good, though it is certainly better than nothing.

(c) Does the simple linear regression model do a good job of predicting the Y values? Make sure you justify your answer.

Solution: This was one of the most frequently missed questions on the exam on which it appeared. In order to tell whether a regression makes good predictions, you need to know how big the errors made by the regression are. One way of evaluating this is to look at the typical distance from the points to the regression line. This number is estimated by $s_{Y|X} = \sqrt{MSE}$. This number can be found as Root MSE on the printout, or by taking the square root of MSE from the ANOVA table. Here $RMSE = \sqrt{123.57} = 11.1161$. To tell whether this means the errors are large, we must compare RMSE to the Y values we are trying to predict. The Y values in this problem range from 298 to 336. Thus we are making an error of roughly 3-4%. This seems pretty good. However, we really should consider the context of the problem. The errors we are making are on the order of 11 inches–nearly a foot. Long jump competitions are usually decided by much less than this so our errors, in context, are still rather large. Note: Many people tried to use R^2 or an F test to say whether the model is a good predictor. These values try to get at whether the model explains a lot of variability. You can explain quite a lot of variability and still have bad predictions.

(d) Is there a significant linear relationship between years and distance? Justify your answer using an appropriate test.

Solution: We could use either a t test or an F test since they are the same for simple linear regression. Our null and alternative hypotheses are

$H_0 : \beta_1 = 0$–i.e. there is not a significant relationship
$H_A : \beta_1 \neq 0$–i.e. there is a significant relationship between the year and the distance of the winning long jump.

From the printout, the test statistics are $t_{obs} = 3.03$ for the t test, and $F = 9.2057$ for the F test. In both cases, the p-value for the test is .0190 which is much less than $\alpha = .05$. Therefore, we reject the null hypothesis and conclude that there is a significant linear relationship between year and the winning long jump distance. To get full credit, you only needed to quote the p-value and explain your conclusions.

(e) In 1968, the Olympics were held in Mexico City, and many records were set, probably due to the high altitude. Explain what diagnostics you could use to determine whether this point is an outlier or an influential point and what each one would tell you. Intuitively do you expect the point to be highly influential? Does it appear to have high leverage? Is it an outlier? Explain. What would happen to your answers to (b)-(d) if this point were removed?

Solution: We could use a whole slew of diagnostics to determine the status of the point: the studentized residual to see if what was an outlier in the sense of a big error, the leverage value to see how much ability it had to tilt the regression line, DFFits, DFBetas and Cook's Distance to see how much effect the point had on the fitted value or regression coefficients. I suspect that 1968 point is both an outlier and an influential point but not a high leverage point. Visually it sticks well up above the path of the rest of the points. Since the data set is small it will probably be influential, shifting the whole line up. However since its X value is right in the middle of the data set it will probably not have super high leverage and the actual slope of the line may not change much. If the point is removed, the regression line will go right through the middle of the rest of the points. Thus the amount of unexplained variability will be smaller and the amount of explained variability will be higher. This will cause R^2 to go up, $s_{Y|X}$ to go down (and hence we will get better predictions), and F to increase (resulting in a lower p-value for our test). It is never easy without removing a point and refitting the regression to tell just how influential that point is. In this case it turns out the point is highly influential. For instance, R^2 goes up from 56% to well over 90% if you refit the model without the point. You can actually try this out from the given data if you want to.

(f) Use STATA to get the residual, histogram, and normal quantile plots for the simple linear regression (or if you don't want to take the time to do so just look at the scatterplot above.) Does it appear that any of our regression assumptions have been violated? Make sure you state each of the assumptions that can be checked with each plot and whether you think they are OK. What do you think is causing any problems you see, and how might you fix them?

Solution: Using a residual plot we can check mean 0, constant variance, and independence/appropriateness of the linear model. Normality can be checked with a histogram and/or QQ plot.

From the residual plot, it appears that the mean 0 assumption is violated. For most values of X, the residuals are all negative. We need the residuals to be centered about the line. This is caused in part by the influential point in 1968. If we took the point out, the regression would go more through the middle of the remaining points and the residuals would be more balanced. However I think there might still be a curved shape, from the dip in the 60s and 70's so I would still consider this assumption to be violated.

Whether you consider the constant variance assumption to be met depends on whether or not you include the 1968 point. If you do, the spread of the residuals is much wider at 1968 than anywhere else. If you leave the point out, there is a fairly even band around the residuals. In general I prefer not to judge a model bad if there is only a single point causing the problems so I would say this assumption is mostly OK.

The assumption that caused the most disagreement was the one involving independence/appropriateness of the linear model. I see a bit of a curved pattern in this data but it is hard to tell, especially given the 1968 point and the fact that we only have one observation every 4 years whetehr this is meaningful. We gave credit on this assumption either way as long as people explained carefully.

Normality also looks a little questionable for this data as the histogram is not too symmetric and the points in the quantile plot don't follow the straight line all that well. It's not terrible but overall I would say this assumption is violated too.

No matter what you think is the right model for this data, the 1968 point makes the error assumptions much more questionable. Removing it will definitely improve the model. It is OK to remove the point in this case because we have a good reason to believe it is abnormal and not representative of what will happen in the future. Mexico City is at an extremely high altitude and this resulted in abnormally strong performances in the short distance track and field events.

(g) A zealous sports fan suggests that the winning distance in the long jump cannot increase for ever, but should instead level off. He therefore suggests fitting a curvilinear regression to the data. The second printout shows the results of fitting the model

$$Y = \beta_0 + \beta_1 Year + \beta_2 Year^2$$

Is it worth adding the term $Year^2$ to the model based on the data presented here? Answer this question using an appropriate test. Make sure you state the null and alternative hypotheses, the p-value for the test, and your conclusions. Is this likely to introduce multicollinearity to the model? Explain why it might, how you could check, and what you could do to fix the problem if it exists. Try it and see if it helps. Finally, in real-world terms is the quadratic model likely to be completely appropriate for this data? Can you suggest an alternate transformation that might be better? Explain.

Solution: First, I certainly agree with the sports fan that the winning long jump distance should level off eventually. The real issues are (a) has that leveling off already begun or is a linear model OK in the range of data we have, and (b) is a parabolic model the right one to take into account the leveling off. The data does not curve too much, so I suspect the answer to (a) is that a linear model is OK for now. Logically, I think the answer to (b) is no–parabolas do not level off as X increases. Thus I suspect ahead of time that the term $Years^2$ is not going to add much to this model. To check this, I need to do a t test to see whether $\beta_2 = 0$. The null and alternative hypotheses are as usual

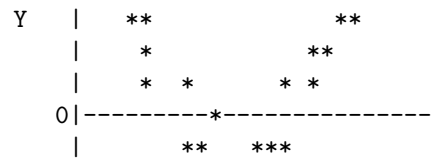
$H_0 : \beta_2 = 0$–i.e. $Years^2$ does not add anything to the model beyond what was already given by Year
$H_A : \beta_2 \neq 0$–i.e. $Year^2$ does make a significant contribution to the model

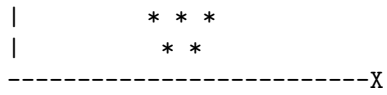
From the curvilinear regression printout, the test statistic is $t_{obs} = -1.59$ and the p-value is .163. Since this p-value is much larger than $\alpha = .05$ we fail to reject the null hypothesis and conclude that $Year^2$ adds nothing new to the model. It is not worth including when Year is already in the model. Note: Many people who took this exam tried to use an F test. This is a multiple regression problem. In multiple regression, an F test checks whether the variables collectively are useful. In this case, the F test is significant. However, that only tells us that at least one of Year and Year^2 is useful–it says nothing specific about Year^2.

There probably is multicollinearity here between X and X^2 as from the plot even if the relationship is curved we are in the part of the parabola where the relationship is not that nonlinear. We could check this by getting either the correlation between X and X^2 or by computing the variance inflation factors. We could center the predictors (by subtracting off the mean of X) to reduce the problem. It is possible this could actually improve the significance of the coefficients by reducing the standard error though I doubt it will help much in this case. A quadratic model is probably not the right choice. Using an inverse model (1/X) might work better as that will actually level off as X gets bigger in line with our physical expectations.

(3) Regression Assumptions

A residual plot from a simple linear regression analysis is shown below. It is followed by four statements about the error assumptions for this model. In each case, say whether the statement is correct. If the statement is not correct, give an appropriate statement about the error assumption referred to.





(a) The mean 0 assumption is correct because there are approximately as many residuals above the line as below it.

Solution: This is **FALSE**. We need the points to be centered about the line for EVERY value of X, not just overall. The mean 0 assumption is clearly violated for this plot. The residuals are positive, then negative, then positive again.

(b) The constant variance assumption is violated because there is a curved pattern to the data.

Solution: This is **FALSE.** The constant variance assumption has nothing to do with whether there is a curved pattern to the data. It has to do with whether the points have the same spread for each value of X. If we draw a band about these points it seems to be of roughly constant width. Thus the constant variance assumption is not violated.

(c) The errors for this data set are approximately normally distributed.

Solution: We can't really tell that from the residual plot. We would need a histogram or a normal quantile plot to determine this properly.

(d) A linear model is not appropriate for this data set because of the curved pattern in the data.

Solution: This is **TRUE.** The curved pattern in the points suggests that a polynomial model is probably more appropriate for this data.