Solutions To Homework Assignment 3

General Comments:

- The solutions given below are (quite a bit) more extensive than would have been necessary to get full credit. I use the answer key as an opportunity to make important points, or mention commonly made mistakes. Nonetheless, the answer key should give you an idea of the type of solutions I would like to receive.
- I have included the graphics in a separate file since they don't import super easily into my mathematical word processing program.

Warmup Problems

(1) Probit Regression Basics:

(a) For a probit model the distribution of Y is assumed to be binomial (or, if you prefer, *Bernoulli* which is a special name for a binomial with a single trial which is what we get for each observation in a logistic model). The link function is the **probit,** Φ which is the c.d.f. or cumulative distrubution function for the standard normal. Basically you can think of the probit function as taking a Z-score and returning the probability of a normal distribution being less than that value, namely $\Phi(z) = P(Z \leq z)$. The systematic component, as usual, is the linear combination of the predictor variables, $\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$. We assme that $p = \Phi(X\beta)$.
(b) Historically the probit model arose out of toxicology or dose response studies. The intuition was that an organism would have a tolerance for a certain dose of a toxin; if the dose the organism was exposed to was below the tolerance then the organism would survive (or not have an adverse reaction) while if the dose received was above the tolerance the organism would die (or have an adverse reaction.) The tolerances were assumed to be normally distributed with means that might depend on certain factors. Thus for an arbitrarily selected organism with a certain set of characteristics the probability of dying/having an adverse reation at a given dose would be a normal probability based on the tolerance distribution for that set of characteristics. This leads naturally to the probit link which effectively takes the tolerance distrubution and converts it to the probability of an adverse reaction. However, it turns out that the probit link is extremely similar to the logistic link and so a probit model can be used in basically any situation where the logistic model is used. The two models produce almost identical results and you can't differentiate easily between them unless either the sample is very large or there are a lot subjects with predicted probabilities in the extreme tails where the distributions are the most different. Because of this the logistic model, which has much more convenient interpretations in terms of odds ratios, has come to dominate the essentially equivalent but harder to interpret probit model. However the probit model is still often used in dose-response studies.

(2) Multinomial and Ordinal Logistic Regression Basics:

(a) For a multinomial logistic regression we assume that the distribution of the outcome, Y, is multinomial with an unknown set of probabilities, $p_0, p_1, p_2, \ldots, p_k$ with $p_0 + p_1 + p + 2 + \cdots + p_k = 1$ for the $k + 1$ different outcome categories. This is just a generalization of the binomial distribution which has two categories, with probabilities that total to 1. We use the same logistic link function as with a standard logistic model (which can just be thought of as a 2-category multinomial model). Our systematic component $\beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$ also has the same form but we need to allow different coefficients for comparing the

odds of each of the possible categories relative to the reference category. In a logistic regression the reference category is "no event" and we have one set of coefficients comparing "event" to "no event". In a multinomial model with k+1 possible categories there will be k sets of β coefficients.

(b) In a logistic regression we defined the odds as $\frac{p}{1-p}$ where p was the probability of the event at a given set of predictor values. In multinomial logistic we pick one of the categories as the reference and calculate "quasi" odds relative to it. For instance, if the 0 category (with probability p_0) is the reference and we want the odds of category j (with probability p_j) relative to that reference then our odds are p_j/p_0. The odds ratio corresponding to a particular variable is obtained, as in logistic regression, by exponentiating the regression coefficients and is interpreted in the same way excepted that the odds in question are relative to the reference category and it is important to specify which category you are comparing to the reference as well as for which variable you are computing an odds ratio.

(c) The multinomial logistic regression described above is completely unconstrained. For each variable you can have an arbitrary odds ratio for comparing each of the categories to the reference category. If the outcome categories are ordered then we need to constrain the model to respect that ordering. The most common way to do this is using something called the proportional odds assumption. Under this assumption, if the outcome Y has possible (ordered) values $0, 1, 2, \ldots, k$ then the ordinal logistic regression model can be formulated as

$$ln(\frac{P(Y \leq j)}{P(Y > j)} = \beta_{j,0} + \beta_1 X_1 + \cdots \beta_m X_m$$

where $j = 0, 1, \ldots, k-1$. Basically we are looking at the odds of **lower values of Y** relative to higher values of Y where the split can be made at any point. The key here is that we assume that the coefficients giving the effect of the X variables are **the same** (and hence the corresponding odds ratios are the same) no matter where we make the split. The only thing that varies depending on the split is the intercept. The result is that the odds ratios are **directional**. We can talk about higher age being associated with better or worse values of the outcome. The proportional odds assumption means that if we look on the log odds scale the relationship between the predictors and the different outcome levels is the same–i.e. we get parallel lines. This is quite a strong assumption. We can relax the assumption by allowing different coefficients at each cut point (generalized ordinal logistitic) or with other groupings of the categories.

(d) Generalized Ordered Logit-Bonus Example: This is a model that people in this class are sometimes interested in so I have included some information about it here although it was not part of the assignment statement and you would not be responsible for it on an exam.

In a generalized ordinal logit model the distribution of the outcome is assumed to be multinomial, the link function is the logit, and the systematic component is a set of linear combinations of the predictors, but we impose constraints on how those linear combinations relate back to the multinomial probabilities. Specifically, the generalized ordered logit amounts to fitting a sequence of separate logistic regressions comparing the odds of higher values of Y to lower values of Y. Its set up of the odds is just like the standard ordinal logistic regression but instead of making the proportional odds assumption that the linear combination is the same for every split (except for the intercept) we allow a different linear combination/set of slope coefficients at every split. Assume Y is coded so that it takes on the values $0, 1, \ldots, k$. Then we fit one logistic model where the event is $Y > 0$ (compared to $Y = 0$), a second model where the event is $Y > 1$ (compared to $Y \leq 1$), a third where the event is $Y > 2$ (compared to $Y \leq 2$) and so on. Specifically, our model is

$$P(Y > j) = logit(X\beta_j) = \frac{e^{\beta_{j,0} + X_1\beta_{j,1} + \cdots X_m\beta_{j,m}}}{1 + e^{\beta_{j,0} + X_1\beta_{j,1} + \cdots X_m\beta_{j,m}}}$$

for $j = 0, 1, \ldots, k-1$. To get the probabilities for the individual levels we note that $P(Y = 0) = 1 - P(Y > 0)$ and $P(Y = j) = P(Y > j-1) - P(Y > j)$ so we can just take successive differences. If you want to read

more about the generalized ordered logit model there's a nice paper in *The Stata Journal* (2006), volume 6 Number 1, pages 58-82.

The generalized ordered logit model is useful when you believe a categorical outcome has a natural ordering but the proportional odds assumption of the standard ordinal logistic model is suspect. For instance, you might have a situation where the spacing between the ordered levels was very uneven in terms of relative severity or where the factor that was most important for moving you from level 0 to 1 was different from the factor most important for moving you from level 1 to level 2. For example, if the outcome categories were no cardiac issues, high blood-pressure and high blood-pressure plus coronary artery disease there is certainly a natural ordering but the difference in severity of illness between the successive levels is not constant. Moreover while the second and third levels would both be affected by factors that contribute to high blood pressure there might be additional factors specific to coronary artery disease that would kick in only in the jump from the second to third level and hence would not follow the proportional odds assumption.

(3) Gasping For Breath Some More:

(a) When the sample is not representative of the rate of cases and controls in the population (e.g. in a case-control study or when one has otherwise over-sampled rare outcomes) then although the slope coefficients and correpsonding odds ratio estimates are valid, the intercept is not and needs to be recalibrated to give corrected predicted probabilities. If $\hat{\beta}_0 = -.047$ is the estimate given by the fitted model, $n_1 = 123$ is the number of cases in the sample, $n_0 = 877$ is the number of negative controls and $P = .05$ is the proportion of cases in the population then the adjusted intercept is

$$\hat{\beta}_0^* = \hat{\beta}_0 + ln(\frac{Pn_0}{(1-P)n_1}) = -.047 + ln(\frac{(.05)(877)}{(.95)(123)}) = -.047 - .980 = -1.027$$

Note that this is negative meaning that our predicted probabilities will be LOWER after we adjust. This is to be expected since the proportion of cases in our sample was about 12%, higher than the fraction there is supposed to be in the true population.

(b) If the data had come from sibling pairs then we would have matching and would need to use a conditional logistic regression model. Taking advantage of the pairing would give us additional power; ignoring it would tend to bias the coefficients towards 0. However it would only be the pairs where the outcome was different that would be informative. If we were doing this manually we would select all these pairs, take the differences in the predictor variables between the cases and non-cases and fit a logistic regression with the outcome set to 1 for all pairs and the intercept forced to 0. Fortunately most computer packages now have a conditional logistic routine so we don't have to go through these contortions.

(c) The printout for the probit model is shown below. On the probit scale we can think of the outcome as being a Z-score which inverting the probit function transforms into a probability of asthma. A higher Z-score implies a higher probability of getting asthma. The coefficient of the family history variable says that the Z-score is .65 units higher for people who have a family history of asthma than for people who don't, all else equal. Remember that this is on a "standard normal" scale so you can conceptualize this as a 2/3rds of a standard deviation change on the Z-score scale–which is quite big. Similarly, the coefficient of the pollution variable is .043 meaning that for every extra thousands particles per cm^3 in pollution the asthma Z-score goes up .043. If we imagine multiplying this by 12 we'd get a change in Z-score of .5 so a 12000 particles per cm^3 in pollution corresponds to a half-standard deviation increase on the Z-score scale. Unfortunately there is no easy translation of these coefficients to the probability scale–there isn't even a nice closed-form formula for the probit function. The effect of a particular change on the Z-score level varies depending on where you are. Think of it the following way. If you start from a Z-score of 0 and move half a standard deviation, the probability goes from .5 (using symmetry of the standard normal) to .69 (using a Z-table or STATA's normal distribution commands). However if you start from a Z-score of 2, which corresponds to a probability of

.977, and increase to 2.5 the corresponding probability goes up to .994, a much smaller increase. There just isn't room for much more change.

```
. probit asthma urban pollution ses breastfed famhist genderp3

Probit regression                                 Number of obs   =       1000
                                                  LR chi2(6)      =      85.38
                                                  Prob > chi2     =     0.0000
Log likelihood =  -330.1698                       Pseudo R2       =     0.1145

------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |  -.0370885   .1600689    -0.23   0.817    -.3508179    .2766408
   pollution |   .0430641   .0108244     3.98   0.000     .0218487    .0642795
         ses |  -.0456595   .0187489    -2.44   0.015    -.0824067   -.0089123
   breastfed |   .0027856   .0055296     0.50   0.614    -.0080522    .0136235
     famhist |   .6547034   .1246158     5.25   0.000     .4104609    .8989458
    genderp3 |  -.6111885   .1147496    -5.33   0.000    -.8360935   -.3862835
       _cons |  -.0704406   .8948964    -0.08   0.937    -1.824405    1.683524
------------------------------------------------------------------------------
```

(d) To get the predicted probability we just find the linear combination for the indicated X values and then apply the inverse probit function. This amounts to looking up the probability corresponding to the linear combination in a Z table. Here the linear combination is

$$Z = -.0704 - .0371(1) + .04306(35) - .04566(50) + .00279(6) + .6547(1) - .61(0) = -0.212$$

The corresponding probability is $P(Z \leq -.212) = .416$ so the boy has a 41.6% chance of developing asthma. The printout for the probability calculation in STATA is shown below. On homework 2 we got a probability of 42% using logistic regression. The difference is essentially due to rounding. Probit and logit models generally produce extremely similar predictions.

```
. display normal(-.212)
.41605352
```

(e) The printout for the logistic model is shown below. The log likelihood is -331.8 compared to -330.2 for the probit model. These values are very similar with the probit value being just slightly better. It is reasonable to compare the the likelihoods since in both cases the distribution for each value is assumed to be biomial with a success probability that depends on the X's. In other words we are maximizing the same likelihood function–it is just a question of which fitted model produces probabilities that match the data better. However we can't exactly do a likelihood ratio chi-squared test to compare the two models as they are not nested in each other. In fact they have exactly the same number of parameters for the same variables, just with slightly different fits. This makes it hard to judge what a big difference in the log likelihoods is though from our experience the observed difference seems pretty small.

```
. logit asthma urban pollution ses breastfed famhistp3 genderp3
```

```
Logistic regression                               Number of obs   =       1000
                                                  LR chi2(6)      =      82.04
                                                  Prob > chi2     =     0.0000
Log likelihood = -331.83839                       Pseudo R2       =     0.1100

------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |  -.0771939   .3020666    -0.26   0.798    -.6692335    .5148458
   pollution |   .0766367   .0200131     3.83   0.000     .0374118    .1158616
         ses |  -.0818089   .0345813    -2.37   0.018     -.149587   -.0140308
   breastfed |    .005728   .0103221     0.55   0.579    -.0145029    .0259589
   famhistp3 |   1.180125   .2230401     5.29   0.000     .7429742    1.617275
    genderp3 |  -1.094722   .2178808    -5.02   0.000    -1.521761   -.6676835
       _cons |  -.0470419   1.655807    -0.03   0.977    -3.292364     3.19828
------------------------------------------------------------------------------
```

(4) The Accidental Statistician:

(a) The printout for the required multinomial logistic regression is shown below. To test whether the model overall is significant we perform a likelihood ratio chi-squared test just as in regular logistic regression. Our null hypothesis is that all the slope coefficients are 0 (so none of the predictors has any effect on any of the levels of the outcome) and our alternative is that at least one of the slope coefficients is non-zero so at least one of the predictors is important for at least one of the levels of the outcome. Here we are asking whether any of the variables fatality, speed, time of day, or gender is related to the type of accident (solo, two car or multi-car). The p-value for the overall likelihood chi-squared test is essentially 0 so we know that at least one of these predictors is useful for distinguishing among the types of accidents.

```
. mlogit type fatal dark speed genderp4

Multinomial logistic regression                   Number of obs   =       300
                                                  LR chi2(8)      =     37.82
                                                  Prob > chi2     =     0.0000
Log likelihood = -284.50935                       Pseudo R2       =     0.0623

------------------------------------------------------------------------------
        type |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
       fatal |   .9899634    .366657     2.70   0.007     .2713289    1.708598
        dark |  -.8833962   .3137976    -2.82   0.005    -1.498428   -.2683642
       speed |  -.0065732   .0060299    -1.09   0.276    -.0183917    .0052452
    genderp4 |    .101886   .2797661     0.36   0.716    -.4464455    .6502174
       _cons |   .2241351   .4382978     0.51   0.609    -.6349128    1.083183
-------------+----------------------------------------------------------------
2            |
       fatal |   1.075279   .6026145     1.78   0.074    -.1058235    2.256382
        dark |  -2.320463   .4875132    -4.76   0.000    -3.275971   -1.364955
       speed |  -.0131464    .007836    -1.68   0.093    -.0285047     .002212
    genderp5 |   -.005185    .367741    -0.01   0.989    -.7259442    .7155742
```

```
       _cons |   .4716146   .5373214     0.88   0.380     -.581516    1.524745
------------------------------------------------------------------------------
(type==0 is the base outcome)
```

(b) The reference category, $Y = 0$ corresponds to 2-car accidents. The outcome $Y = 1$ is solo accidents and the outcome $Y = 2$ is 3 or more car accidents. The coefficient for the fatal indicator in the $Y = 1$ table is $b_1 = .99$ which means that all else equal the log odds comparing the likelihood of a solo accident to a 2-car accident is .99 higher if the accident involves fatalities than if it does not. While technically correct this is not very enlightening! Practically the implication is that solo accidents are more likely to involve fatalities than two-car crashes (perhaps because solo accidents are less likely unless a person is driving quite recklessly). If we convert this to an odds ratio, $e^{.99} = 2.69$ so the odds of an accident being solo (as compared to a 2-car accident) are over 2.5 times as high if the accident involved a fatality than if it did not. Similarly, for fatal accidents the log odds of a 3-or-more-car accident (relative to a 2-car accident) are 1.08 units higher when fatalities are involved than when they are not, all else equal. The corresponding odds ratio is $e^{1.08} = 2.94$. The odds that the accident involves 3+ cars is three times as high if the accident involves fatalities than if it does not.

The speed variable is continuous so here we are talking about the effects of going 1 mile per hour faster. When we are comparing solo accidents to 2-car accidents, a 1 mph increase in speed is associated with $b_3 = -.006$ or a decrease of .006 in the log odds of a solo accident. The corresponding odds ratio is $e^{-.006} = .99$ or there is a 1% decrease in the odds of a solo accident for each extra mile per hour of speed. This suggests that higher speed makes it less likely the accident is solo but the reduction is very small and not at all close to significant so we can't be sure that speed provides any information about the relative likelihood of solo and 2-car crashes. When we compare 3+ car crashes to 2-car crashes our coefficient is $b_3 = -.013$ and the correpsonding odds ratio is $e^{-.013} = .987$ meaning there is a 1.3% reduction in odds of a 3+ car crash per extra mile per hour of speed. Once again this effect is not very large in magnitude and is not statistically significant so we shouldn't read too much into it.

(c) For distinguishing solo accidents from two car accidents it looks like whether or not the crash involves a fatality (p-value = .007) and whether the accident occurs at night (p-value = .005) are the significant predictors. For distinguishing 3+ car crashes from 2-car crashes it looks as if whether or not the accident occurs at night is the only fully significant predictor although fatality and speed are both close. Fatalities appear to be more highly associated with solo and 3+ car accidents, whereas occurance at night is associated with a higher likelihood of being a 2-car crash. For these individual tests the hypotheses would be written as

$H_0 : \beta_{1,1} = 0$ whether or not the accident involves a fatality has no effect of the odds of being a solo crash (as compared to a 2-car crash) after accounting for time of day, speed and gender.
$H_A : \beta 1, 1 \neq 0$–the odds of being a solo crash are different depending on whether or not the accident involves a fatality, even after adjusting for time of day, speed and gender.

The test statistic for these hypotheses is a Wald Z-statistic.

If we want to test **overall** whether one of the variables is useful we can either perform a likelihood chi-squared test, fitting the model with and without the variable, or we can perform a Wald-type test. The printouts corresponding to these tests are shown below. An example of the hypotheses is given for the time of day variable:

$H_0 : \beta_{1,2} = \beta_{2,2} = 0$–time of day is not associated with type of accident after adjusting for fatalitiy, speed and gender
H_A : at least one of $\beta_{1,2}, \beta_{2,2} \neq 0$–the relative likelihoods of the different accident types does depend on time of day.

From the printouts below we see that overall whether or not the accident is a fatality and what time of day it occurred are significant predictors of the type of accident. However speed and gender are not.

```
(1) Fatality:

(a) LR test version
. mlogit type dark speed genderp4
. estimates store nofatality
. lrtest crashfull nofatality

Likelihood-ratio test                                 LR chi2(2)  =      8.75
(Assumption: nofatality nested in crashfull)          Prob > chi2 =    0.0126

(b) Wald test version
. test fatal

 ( 1)  [1]fatal = 0
 ( 2)  [2]fatal = 0

           chi2(  2) =    8.45
         Prob > chi2 =    0.0146
(2) Time of Day:

(a) LR test version

. lrtest crashfull nodark

Likelihood-ratio test                                 LR chi2(2)  =     31.80
(Assumption: nodark nested in crashfull)              Prob > chi2 =    0.0000

(b) Wald test version

. test dark

 ( 1)  [1]dark = 0
 ( 2)  [2]dark = 0

           chi2(  2) =   25.24
         Prob > chi2 =    0.0000

(3) Speed:

(a)LR test version

. lrtest crashfull nospeed

Likelihood-ratio test                                 LR chi2(2)  =      3.14
(Assumption: nospeed nested in crashfull)             Prob > chi2 =    0.2082

(b) Wald test version
```

```
 . test speed

 ( 1)  [1]speed = 0
 ( 2)  [2]speed = 0

           chi2(  2) =    3.09
         Prob > chi2 =    0.2128
(4) Gender:

(a) LR test version

. lrtest crashfull nogender

Likelihood-ratio test                                     LR chi2(2)  =      0.15
(Assumption: nogender nested in crashfull)                Prob > chi2 =    0.9258

(b) Wald test version

. test genderp4

 ( 1)  [1]genderp4 = 0
 ( 2)  [2]genderp4 = 0

           chi2(  2) =    0.15
         Prob > chi2 =    0.9256
```

(d) We need to look at the fatality and time of day variables. We already know we have no evidence of a difference across levels for speed and gender because we can't even reject the idea that any of the coefficients are different from 0. To test whether the coefficients, for instance for fatality, are the same our hypotheses are

$H_0 : \beta_{1,1} = \beta_{2,1}$–whether or not there is a fatality has the same impact on the odds of a solo accident as on the odds of a 3+-car accident (relative to a 2-car accident) after adjusting for gender, speed and time of day. $H_A : \beta_{1,1} \neq \beta_{2,1}$–the association between whether or not there is a fatality and type of accident is different depending on whether you are talking about solo vs 2-car accidents or 3+ vs 2-car accidents.

These hypotheses are usually tested using the Wald procedure. The printouts are given below. We see that there is no evidence of a difference in the effect of fatality across the various levels (p-value .89). However the test for the time of day variable is significant (p-value .0044). Thus we see that the effect of time of day is different depending on whether we are talking about the odds of a solo creash or the odds of a milti-car crash.

```
. test [1=2]: fatal

 ( 1)  [1]fatal - [2]fatal = 0

           chi2(  1) =    0.02
         Prob > chi2 =    0.8892

. test [1=2]: dark

 ( 1)  [1]dark - [2]dark = 0
```

```
  chi2(  1) =    8.10
Prob > chi2 =    0.0044
```

(e) Overall it seems that whether or not a fatality is involved and when the accident occurred are associated with how many cars are involved in the accident, but speed and gender do not differ across accident types. Specifically, it seems that fatalities are associated with higher odds of solo or multi-car crashes relative to two-car crashes, the with the increase in odds being similar for the two types (although we are more sure it is significant for solo crashes than for multi-car crashes. (This may be because there were fewer multi-car crashes in our data set so our power for this comparison is lower.) On the other hand, driving after dark was associated with lower odds of solo or multi-car crashes, with the effect being stronger for multi-car crashes. This makes some intuitive sense. There is less traffic at night so there is less likely to be a many-car pile-up.

(f) To get the predicted probabilities we use the equations

$$P(Y=0) = \frac{1}{1+e^{X\beta_1}+e^{X\beta_2}}$$

$$P(Y=1) = \frac{e^{X\beta_1}}{1+e^{X\beta_1}+e^{X\beta_2}}$$

$$P(Y=2) = \frac{e^{X\beta_2}}{1+e^{X\beta_1}+e^{X\beta_2}}$$

The easiest thing is to compute the systematic components and then plug them in to the various equations. For a man driving 90 miles per hour after dark with no fatalities we have $X_1 = 0, X_2 = 1, X_3 = 90, X_4 = 0$. Thus our systematic pieces are

$$X\beta_1 = .224 + .99(0) - .883(1) - .0066(90) + .101(0) = -1.199$$

$$X\beta_2 = .472 + 1.075(0) - 2.32(1) - .0131(90) - .005(0) = -3.027$$

The corresponding predicted probabilities are

$$P(Y=0) = \frac{1}{1+e^{-1.199}+e^{-3.027}} = .74$$

$$P(Y=1) = \frac{e^{-1.199}}{1+e^{-1.199}+e^{-3.027}} = .22$$

$$P(Y=2) = \frac{e^{-3.027}}{1+e^{-1.199}+e^{-3.027}} = .04$$

It appears that the chances of such a person being involved in a 2-car accident are 74%, the chances of being involved in a solo accident are 22% and the chances of being involved in a multi-car accident are 4%.

(g) Now we are asked to fit an ordinal logistic regression. The printout below correponds the proportional odds logistic regression model. The overall likelihood ratio chi-squared test has a p-value of 0, meaning that at least one of fatality, time of day, speed and gender is useful for discriminating among the accident types. This is hardly a surprise given the results of parts (a)-(f)! The hypotheses in this case are

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$–none of the predictors is useful
H_A : At least one $\beta_j \neq 0$–at least one of the predictors is associated with accident type.

Note that we only have one coefficient for each variable now so it is less messy to write down the hypotheses!

```
. ologit type fatal dark speed genderp4

Ordered logistic regression                       Number of obs   =        300
                                                  LR chi2(4)      =      31.08
                                                  Prob > chi2     =     0.0000
Log likelihood = -287.88067                       Pseudo R2       =     0.0512

------------------------------------------------------------------------------
        type |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       fatal |   .9140756   .3251895     2.81   0.005     .2767158    1.551435
        dark |  -1.467922   .2768451    -5.30   0.000    -2.010528   -.9253156
       speed |  -.0090571   .0051082    -1.77   0.076    -.0190689    .0009548
    genderp4 |   .0525099     .23924     0.22   0.826    -.4163918    .5214116
-------------+----------------------------------------------------------------
       /cut1 |  -1.126108   .3723292                     -1.85586   -.3963565
       /cut2 |    .608556   .3693093                     -.115277    1.332389
------------------------------------------------------------------------------
```

(h) In an ordinal logistic regression the coefficients and odds ratios for the X variables tell you how the predictors are associated with the odds of **lower** vs **higher** values of Y and the intercepts calibrate the respective probabilities of the different categories. The traditional model gives odds for the lower categories of Y compared to the higher catgories. However STATA reverses the signs on its coefficients so that a positive value of a coefficient or an odds ratio higher than 1 corresponds to a higher probability of a higher value of Y. Here we are thinking of accidents being ordered from least serious ($Y = 0$) to most "serious" ($Y = 2$). Thus the coefficient of fatality tells us that the log odds of a more serious accident are .91 higher if the accident involves fatalities than if it does not (which seems obvious!). The corresponding odds ratio is $e^{.91} = 2.5$ which means the odds of a more serious accident are 2.5 times as high if there is a fatality than if there is not, all else equal. The coefficient of the dark variable tells us that the log odds of a more serious accident go down 1.47 points if the accident takes place at night, all else equal. The corresponding odds ratio is $e^{-1.46} = .23$. In other words there is a 77% reduction in odds of a more serious accident at night compared to during the day. Perhaps this is because there is less traffic at night. The coefficient for the speed variable tells us that for every extra mile per hour the log odds of a more serious accident goes down .009 units. The corresponding odds ratio is $e^{-.009} = .991$ meaning there is less than a 1% reduction in odds of a more serious accident for each mile per hour faster that you drive. Here the negative sign seems very counter-intuitive–one would expect that faster driving leads to worse accidents. However it is possible that people drive faster during the day and that fatalities are also associated with faster driving which may have already accounted for the effects of speed. We could check for a multicollinearity problem by looking at the relationships among the assorted variables. The odds ratio for the gender variable is positive, suggesting that females have a higher odds/probability of being involved in more serious accidents. This also runs counter to conventional wisdom. However the coefficient is highly non-significant so we really shouldn't be trying to make such an interpretation. Finally we should note that my "ordering" of the accident types is a little artificial which may explain some of the results!

(i) Now we need to get the predicted probabilities for the ordinal logistic regression. The way the ordinal logistic model is set up we have

$$ln(\frac{P(Y=0)}{P(Y\neq 0)}) = \beta_{0,0} + \beta_1 X_1 + \cdots$$

In analogy with logistic regression it follows that

$$p_0 = P(Y = 0) = \frac{e^{X\beta_0}}{1 + e^{X\beta_0}}$$

where by $X\beta_0$ I mean the expression on the right side of the first equation above. Similarly we have

$$ln(\frac{P(Y = 0, 1)}{P(Y \neq 0, 1)}) = \beta_{1,0} + \beta_1 X_1 + \cdots$$

To get $p_1 = P(Y = 1)$ we just need to subtract the probability that $Y = 0$ from the probability that $Y = 0$ or 1. We continue this way sequentially. The final thing we need to keep in mind is how our package reports the assorted coefficients. For example in STATA I noted that the coefficients reported correspond to the log odds of **higher** values of Y. Thus for the probability calculations above which focus on getting the lower values of Y we need to take negatives. STATA's cut points correspond to the constant terms. Thus for predicting the probability of a two-car accident with no fatalities ($X_1 = 0$), at night ($X_2 = 1$), at 90 miles per hour ($X_3 = 90$) for a man ($X_4 = 0$) we use the following linear combination of the X values, remembering to reverse the signs on our coefficients. Our value for $\hat{\beta}_{0,0}$ is what STATA calls cut1:

$$-1.126 - .91(0) + 1.46(1) + .009(90) - .052(0) = 1.144$$

The corresponding predicted probability is

$$\frac{e^{1.144}}{1 + e^{1.144}} = .758$$

In a similar manner we get the linear combination for the probability that $Y = 0$ or $Y = 1$ using cut 2 as

$$.609 - .91(0) + 1.46(1) + .009(90) - .052(0) = 2.879$$

The corresponding probability is

$$\frac{e^{2.879}}{1 + e^{2.879}} = .947$$

It follows that $p_1 = .947 - .758 = .189$. Finally, since we only have three categories we can get $p_2 = 1 - p_0 - p_1 = 1 - .947 = .053$. Thus it seems a night-time accident involving a man driving 90 miles per hour with no fatalities has a 75.8% chance of being a 2-car accident, an 18.9% chance of being a solo accident and a 5.3% chance of being a multi-car accident.

(j) Our results with the ordinal logistic regression seem fairly similar to those from the multinomial logit. The most significant predictors are fatalities and time of day, with fatalities being associated with "more serious" (i.e. solo or multi-car accidents) and night time being associated with "less serious" accidents. The main difference is that the speed variable seems closer to significance in the ordinal logistic model than the multinomial model though even here it doesn't meet the .05 cutoff. The predicted probabilities we got for our test case were also very similar in the two models. One way we can formally compare the models is to look at the log likelihoods. The log likelihood for the multinomial model is -284.50935 while that for the ordinal logistic model is -287.88067. Although the models are not exactly nested -2 times the difference in log likelihoods is still fairly close to having a chi-squared distribution with degrees of freedom equal to the difference in number of model parameters. Here this difference would be 6.74 and there are four extra parameters in the multinomial model, one for each predictor variable. We can use the chi2tail command from homework 2 to get the associated p-value as follows:

```
. display chi2tail(4,6.74)
.15028272
```

This p-value is fairly large so it doesn't look like the more flexible multinomial model is a significant improvement over the simpler ordinal logistic model and we can safely use the results from the parts (g)-(i).

Turn-In Problems

(5) Cancer Conundrums:

(a) The conditional logistic regression printout is shown below. Since exposure is the only variable in the model it is easy to evaluate its significance using either the overall likleihood ratio chi-squared test or the Wald test. Using the likleihood ratio test the p-value is 0 to four decimal places while with Wald test the p-value is .001. (As usual the Wald test is a bit more conservative.) Either way we see that exposure is a highly significant predictor of whether or not someone gets cancer. The corresponding odds ratio is obtained by exponentiating the regression coefficient. We have $OR = e^{.3319} = 1.39$. This tells us that on average the odds of getting cancer go up 40% for each extra point of chemical exposure. This does not sound good, particularly when we realize the exposure variable ranged from 0 to 34 so a 1 point increase is not especially unusual! Of course the base rate of cancer may have been quite low.....

```
. clogit cancer exposure, group(pairid)

Conditional (fixed-effects) logistic regression   Number of obs   =        100
                                                  LR chi2(1)      =      25.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -21.966428                       Pseudo R2       =     0.3662

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |   .3319311   .0961554     3.45   0.001     .1434699    .5203923
------------------------------------------------------------------------------
```

(b) The printout for the standard logistic model ignoring the pairing is shown below. I used the logistic printout to get the odds ratio automatically. The odds ratio is considerably lower at 1.10–we now only get an estimated 10% increase in odds of cancer per additional point of exposure–only a quarter as large an effect. The p-value is also less significant (.0016 by the likelihood ratio chi-squared test or .003 by the Wald test) although exposure is still clearly an important predictor. This is consistent with what I would have expected. Taking advantage of the pairing provides extra power and should lead to increased significance. Ignoring the pairing when it is there tends to bias the coefficients towards 0 or odds ratios towards 1.

```
. logit cancer exposure

Logistic regression                               Number of obs   =        100
                                                  LR chi2(1)      =      10.01
                                                  Prob > chi2     =     0.0016
```

```
Log likelihood = -64.310035                     Pseudo R2       =     0.0722

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |   .0996794   .0340266     2.93   0.003     .0329884    .1663704
       _cons |  -1.427561   .5257191    -2.72   0.007    -2.457951   -.3971706
------------------------------------------------------------------------------
*******************************************************************************
. logistic cancer exposure

Logistic regression                             Number of obs   =        100
                                                LR chi2(1)      =      10.01
                                                Prob > chi2     =     0.0016
Log likelihood = -64.310035                     Pseudo R2       =     0.0722

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |   1.104817   .0375932     2.93   0.003     1.033539     1.18101
------------------------------------------------------------------------------
```

(c) When the sample is not representative of the rate of cases and controls in the population (e.g. in a case-control study or when one has otherwise over-sampled rare outcomes) then although the slope coefficients and correpsonding odds ratio estimates are valid, the intercept is not and needs to be recalibrated to give corrected predicted probabilities. If $\hat{\beta}_0$ is the estimate given by the fitted model, n_1 is the number of cases in the sample, n_0 is the number of non-cases in the sample and P is the proportion of cases in the population then the adjusted intercept is

$$\hat{\beta}_0^* = \hat{\beta}_0 + ln(\frac{Pn_0}{(1-P)n_1})$$

In this case we have an equal number of cases and controls, $n_0 = n_1 = 50$ so those terms cancel out. We are told the true fraction in the population is around $P = .05$. Thus our adjusted intercept is

$$\hat{\beta}_0^* = -1.428 + ln(\frac{(.05)}{(.95)}) = -1.428 - 2.944 = -4.372$$

Since this value is very negative our predicted probabilities will be much lower after the adjustment which is logisical since we have way over-sampled cases in the study relative to the population prevalence. We are also asked to check how much this affects the predicted probability of cancer for a person with an exposure level of 20. Using the unadjusted model the predicted probability is

$$p_U = \frac{e^{-1.428+.0997(20)}}{1+e^{-1.428+.0997(20)}} = .638 = 63.8\%$$

For the adjusted model we have

$$p_A = \frac{e^{-4,372+.0997(20)}}{1+e^{-4.372+.0997(20)}} = .085 = 8.5\%$$

The difference in the predicted probabilities is enormous. Again, this is not surprising since in our sample our rate of cases was 10 times as large as in the actual population.

(6) I Await Your Response:

(a) The prinout for the probit model is shown below. Looking either at the overall likelihood ratio chi-squared test or the Wald test for the dose variable we have a p-value of essentially 0, meaning that there is a significant relationship between response and dose. Since the coefficient of the dose variable is positive the implication is that the chances the patients respond to the medication increases as the dose gets higher which is hardly a surprise. Probit models were developed in the context of toxicity or dose response studies and are often still used in this context. The intuition is that there is an underlying (unobserved) tolerance score such that if the dose given exceeds the tolerance score then you get a response and if not you don't. If the tolerance scores are normally distributed in the population (either overall or within subsets specified by model covariates) then the probit link which is the inverse normal distribution makes sense. However as we will see below, the probit and logit models usually produce very similar responses and the logit is easier to interpret so probit models have become less common.

```
. probit response dose

Probit regression                                 Number of obs   =        500
                                                  LR chi2(1)      =      82.10
                                                  Prob > chi2     =     0.0000
Log likelihood = -304.36807                       Pseudo R2       =     0.1188

------------------------------------------------------------------------------
    response |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .0227134   .0025996     8.74   0.000     .0176183    .0278085
       _cons |  -.8093618   .1189914    -6.80   0.000    -1.042581   -.5761429
------------------------------------------------------------------------------
```

(b) In a probit regression we are interested in the probability of an event but we do the modeling on the probit scale where it is reasonable to model the outcome as a linear combination of the predictors. The intercept in this model is the probit score when the predictors are all 0. Here that means that the intercept is the probit score when the dose is 0–i.e. the subject has received no medication. If we transform back we will get the probability of a response for a person who has not received treatment. You can think of this as the spontaneous recovery rate or the placebo response rate if people in the study are randomized to receive medication or placebo. To get the probability we need to apply the inverse logit function which amounts to taking the predicted probit score and looking it up on a Z table to find the probability of being less than that value. Here we need

$$P(Z \leq -.8094) = .209$$

You could do this from a Z table but I used STATA's normal probability calculator:

```
. display normal(-.80936)
.20915405
```

There is about a 20.9% spontaneous or placebo response rate predicted by our model.

(c) Since the normal distribution is symmetric a Z-score of 0 corresponds to a probability of 50% ($P(Z \leq 0) = .5$). Thus to fund the dose at which the response rate is 50% we simply need to find the dose that yields a probit score of 0. We therefore need

$$0 = b_0 + b1_dose = -.809 + .0227dose$$

or

$$dose = \frac{.809}{.0227} = 35.6$$

We need a dose of about 35.6 in whatever units this medication is administered. (The mean dose in our sample was about 39 so this doesn't seem inconsistent.)

(d) The Hosmer-Lemeshow test works exactly the same way for probit regression as for logistic regression– we are modeling the same outcome with the same likelihood function–all that has changed (very slightly) is the link function used in the model. The printout is below. The p-value for the H-L test is hugely significant (p-value essentially 0) which means there is a substantial discrepancy between our model and the observed response pattern in our data. Our model is NOT well calibrated. We suspect this is because a linear relationship is not appropriate, at least not forever. Eventually as the dose gets higher the response rate will not continue to increase and may in fact even go down if too high doses are toxic. We could check this directly by plotting the actual counts and the predicted counts (sums of probabilities) for different dose bins as was done on HW3 and see if they diverged as expected at the edges of the dose distribution. I have included the table option as part of my test so that we can see the observed and expected counts. They do bear out this idea. The observed counts of response (Obs_1)are much less than the predicted counts (Exp_1) for high and low dose groups and much higher than the predicted counts in the middle which suggests a quadratic pattern.

```
. estat gof, group(10) table

Probit model for response, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
  +--------------------------------------------------------+
  | Group |   Prob | Obs_1 | Exp_1 | Obs_0 | Exp_0 | Total |
  |-------+--------+-------+-------+-------+-------+-------|
  |     1 | 0.2541 |     1 |  11.6 |    49 |  38.4 |    50 |
  |     2 | 0.3197 |     7 |  15.8 |    47 |  38.2 |    54 |
  |     3 | 0.3871 |    21 |  17.1 |    27 |  30.9 |    48 |
  |     4 | 0.4581 |    29 |  22.0 |    22 |  29.0 |    51 |
  |     5 | 0.5169 |    31 |  23.0 |    16 |  24.0 |    47 |
  |-------+--------+-------+-------+-------+-------+-------|
  |     6 | 0.5931 |    43 |  32.4 |    15 |  25.6 |    58 |
  |     7 | 0.6700 |    36 |  28.1 |     8 |  15.9 |    44 |
  |     8 | 0.7548 |    41 |  40.5 |    15 |  15.5 |    56 |
  |     9 | 0.8020 |    31 |  34.4 |    13 |   9.6 |    44 |
  |    10 | 0.8432 |    27 |  39.5 |    21 |   8.5 |    48 |
  +--------------------------------------------------------+

      number of observations =       500
            number of groups =        10
    Hosmer-Lemeshow chi2(8) =        68.37
                Prob > chi2 =         0.0000
```

(e) The model with dose and dose squared as predictors is shown below. To check whether this model is an improvement over the linear dose model we can use a likelihood ratio chi-squared test to compare the two models or we can simply test directly whether the coefficient of the dose squared variable is 0. The p-value

for the Wald test for the dose squared term is 0, implying that the quadratic term was worth adding to the model. To get the likelihood ratio chi-squared test we would compute -2 times the difference in log likelihoods which is $-2(-204.4-(-268.1)) = 72.6$ which is huge. We don't even need to compute the p-value to know this is significant–the quadratic model is much better than the linear model. We note that the coefficient of the dose squared variable is negative which means on the probit scale our model is a downwards opening parabola. For a while as dose increase the probit score and hence the probability of response increases, but eventually as the dose gets too high the probit score and hence the probability of response begin to decrease again. This is exactly in line with the researchers' theory. To tell if this model is well calibrated we once again use a Hosmer-Lemeshow test. The p-value on the printout below is now .8723, not even close to significant, implying that our model is now properly calibrated. This good since the quadratic model is in fact the one I used the generate the data!

```
. probit response dose dosesq


Probit regression                                 Number of obs   =        500
                                                  LR chi2(2)      =     154.59
                                                  Prob > chi2     =     0.0000
Log likelihood =  -268.1228                       Pseudo R2       =     0.2238

------------------------------------------------------------------------------
    response |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .1209652   .0132557     9.13   0.000     .0949845    .1469459
      dosesq |  -.0011576   .0001468    -7.89   0.000    -.0014453   -.0008699
       _cons |  -2.314196   .2556639    -9.05   0.000    -2.815288   -1.813104
------------------------------------------------------------------------------

 estat gof, group(10)

Probit model for response, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       500
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =      3.83
                 Prob > chi2 =      0.8723
```

(f) Now we are asked to identify the optimal dose, which is the one with the highest response rate. The response rate will peak where the probit score peaks so we just have to find the dose associated with the maximum value of our regression equation on the probit scale: $-2.31+.12dose-.0012dose^2$. As noted in the assignment, the peak of a parabola can be determined from the coefficients. Specifically, the dose associated with the maximum score is

$$dose = \frac{-b_1}{2b_2} = \frac{-.12}{2(-.0012)} = 52.25$$

Thus our optimal dose is slightly over 50 in whatever the drug units are. (When I generated the data I used an optimal dose of 50 but of course you get a certain amount of noise with a finite sample!)

(g) We are asked to obtain the fitted probabilities for the probit model which I will call "probitprobs" and also to fit the logistic model and get its predicted probabilities which I will call "logitprobs". The commands for this and the output are shown below. We are then asked to use these values to compare the probit and logistic models. First we need to compare the predicted probabilities when the person receives no medication. For the probit model we get this by finding the probability associated with the intercept value which is

$$P(Z < -2.31) = .0103$$

or about 1%. Note that I needed to redo this using the quadratic model–the value I computed in part (b) was for the inferior linear model. This new value actually makes much more sense since we would expect relatively few people to respond without intervention. For the logistic model we have to use our usual probability formula, again applied to the intercept when the dose is 0:

$$\frac{e^{-3.84}}{1 + e^{-3.84}} = .02$$

This number is very similar to the probit model though not absolutely identical–the two models do differ a bit in the tails. Next we are asked to compare the optimal doses. We found the optimal dose based on the probit model to be 52.25. For the logsitic model we apply the same procedure using the new set of coefficients. The projected peak of the parabola is at

$$dose = \frac{-(.20)}{-2 * (-.0019)} = 52.18$$

This is again extremely similar to what we got from the probit model.

Next we are asked to compare the log likelihoods for the same model. We can't actually do a test since the models are not nested–in fact they have identical predictors, just slightly different estimates for them. However we are assuming the same outcome distribution and the same model for the probabilities so the likelihood function is the same–we are just checking which values of the parameters in that liklekihood function produces a better overall fit. The likelihood for the probit model was -268.1228 and the likleihood for the logit model is -268.83632 , again extraordinarily similar–if we were in a situation where performing a chi-squared test this difference is not one that would be at all significant. While the distribution isn't really right for the test the rough intuition it gives about the magnitude of the difference is.

Next we are asked to compute the correlation between the predicted probabilities. The higher it is the more similar they are. In point of fact we could actually run a simple linear regression and hope for an itnercept of 0 and a slope of 1, suggesting the probabilities were THE SAME–all the correlation does is check whether they are linearly related which is a weaker statement. I have included both printouts below. The correlation is a massive $r = .9999$ and the simple linear regression is very close to the anticipated intercept and slope. Finally we are asked to obtain a scatterplot showing the relationship between the two sets of predicted probabilities. From the correlation and simple linear regression we already know it will look like an almost perfect straight line. The plot is in the accompanying graphics file.

Overall we see that the probit and logistic model have produced extremely similar results although there are some slight differences at the extreme dose values.

```
. predict probitprobs
(option p assumed; Pr(response))

. display normal(-2.314196)
.01032849
```

```
*************************************************************************
. logit response dose dosesq

Logistic regression                               Number of obs   =        500
                                                  LR chi2(2)      =     153.16
                                                  Prob > chi2     =     0.0000
Log likelihood = -268.83632                       Pseudo R2       =     0.2217

------------------------------------------------------------------------------
    response |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .2005425   .0232419     8.63   0.000     .1549891    .2460959
      dosesq |  -.0019218   .0002533    -7.59   0.000    -.0024183   -.0014253
       _cons |  -3.836641   .4538644    -8.45   0.000    -4.726199   -2.947083
------------------------------------------------------------------------------

. predict logitprobs
(option p assumed; Pr(response))
*************************************************************************
. corr probitprobs logitprobs
(obs=500)

             | probit~s logitp~s
-------------+------------------
 probitprobs |   1.0000
  logitprobs |   0.9999   1.0000


. regress probitprobs logitprobs

      Source |       SS       df       MS              Number of obs =     500
-------------+------------------------------           F(  1,   498) =       .
       Model |  34.5337645     1  34.5337645           Prob > F      =  0.0000
    Residual |  .007621841   498  .000015305           R-squared     =  0.9998
-------------+------------------------------           Adj R-squared =  0.9998
       Total |  34.5413863   499  .069221215           Root MSE      =  .00391

------------------------------------------------------------------------------
 probitprobs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  logitprobs |   1.006557   .0006701  1502.13   0.000     1.005241    1.007874
       _cons |  -.0055198   .0003983   -13.86   0.000    -.0063024   -.0047373
------------------------------------------------------------------------------
```

(7) Healthy and Happy?

(a) The printout for the required multinomial logistic model is shown below. We had scored excellent health as 0 and STATA defaults to the 0 (or lowest) category as the reference so we didn't have to coerce this. If the health variable hadn't been scored this way we could either have rescored it or used the **baseoutcome** or **ref** options. As with standard logistic regression we use the overall likelihood ratio chi-squared test to tell us whether the model is significant. Here the test statistic is $\chi^2 = 48.24$ and the correpsonding p-value is

essentially 0. We conclude that at least one of age, income and depression is related to health status (hardly a surprise.)

```
. mlogit health age income depressed

Multinomial logistic regression                 Number of obs   =        294
                                                LR chi2(9)      =      48.24
                                                Prob > chi2     =     0.0000
Log likelihood = -307.02012                     Pseudo R2       =     0.0728

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
         age |   .0281627   .0080175     3.51   0.000     .0124487    .0438766
      income |  -.0177061   .0090582    -1.95   0.051    -.0354598    .0000476
   depressed |   .2772034   .3821483     0.73   0.468    -.4717936      1.0262
       _cons |  -.9915663   .4479902    -2.21   0.027    -1.869611   -.1135216
-------------+----------------------------------------------------------------
2            |
         age |    .038071     .01152     3.30   0.001     .0154923    .0606498
      income |  -.0339626   .0167442    -2.03   0.043    -.0667807   -.0011444
   depressed |   1.083377    .486915     2.22   0.026     .1290413    2.037713
       _cons |  -2.567521   .7215965    -3.56   0.000    -3.981824   -1.153217
-------------+----------------------------------------------------------------
3            |
         age |   .0753349   .0190348     3.96   0.000     .0380273    .1126425
      income |   .0026705   .0213375     0.13   0.900    -.0391501    .0444912
   depressed |   1.793727   .6755939     2.66   0.008     .4695872    3.117867
       _cons |  -6.367643   1.370215    -4.65   0.000    -9.053215   -3.682071
------------------------------------------------------------------------------
(health==0 is the base outcome)
```

(b) For a continuous variable the coefficient gives the change in log odds (of being in the indicated category as compared to the reference category) associated with a 1 unit increase in X, all else held fixed. For an indicator variable the coefficient gives the difference in log odds between subjects who do and do not have the characteristic represented by the indicator. Here we see that each additional year of age is associated with, respectively an increase of .028 in the log odds of good health, an increase of .038 in the log odds of fair health and an increase of .075 in the log odds of poor health, all relative to excellent health. Note that all of these changes are in the direction of increased age being associated with increased chances of poorer health and the magnitudes are getting greater as the degree of poor health gets worse which makes perfect sense. Similarly each extra $1000 of income is associated with a decrease of .018 in the log odds of good health, a decrease of .034 in the log odds of fair health and an increase of .003 in the log odds of poor health, all relative to excellent health. These first two make sense–it seems that higher income is associated with less chance of poorer health. However the final comparison is a little odd–we should note though that it's p-value is highly non-significant so we probably can't read too much into it. Finally, for the depression indicator we note that people who are depressed have a .28 higher log odds of fair health than not depressed people, people who are depressed have a 1.08 higher log odds of fair health than not depressed people and people who are depressed have a 1.79 higher log odds of poor health than not depressed people, all relative to excellent heatlh. This makes sense too–depression is associated with higher odds of poorer health and the

degree of difference is more pronounced the worse the state of health being considered.

These numbers on the log odds scale are, of course, hard to interpret except in terms of their direction and relative magnitude. If we put things on the odds ratio scale it is a little easier. For instance, for age we have odds ratios of $e^{.028} = 1.028, e^{.038} = 1.039$, and $e^{.075} = 1.078$ respectively. This means that each extra year of age is associated with a 2.8% increase in the odds of good health, a 3.9% increase in the odds of fair health and a 7.8% increase in the odds of poor health, all odds calculated relative to excellent health. For income the odds ratios are .98, .97 and 1.003 respectively meaning each extra $1000 of income is associated with a 2% decrease in the odds of good health, a 3% decrease in the odds of fair heatlh and a .3% incerase in the odds of poor health relative to excellent health, all else equal. Finally, for depression the odds ratios are 1.32, 2.94 and 6.01 respectively meaning a depressed person has 32% higher odds of good heatlh, 2.94 times as high odds of fair health and 6 times as high odds of poor health (all relative to excellent health) as a not depressed person, all else equal.

We are also asked for the effect of a 10 year increase in age on the odds of poor health and a $5000 increase in income on the odds of fair health. To get the change in log odds associated with a change of Δ in one of the X variables we simply multiply the relevant coefficient by Δ. To get the odds ratio associated with a Δ unit change we can either multiply the associated coefficient by Δ and then exponentiate or else we can raise the odds ratio for the one-unit change to the power Δ. For poor health the coefficient of age is $b_{3,3} = .075$ so the change in log odds associated with a 10 year increase in age is $10b_{3,3} = 10(.075) = .75$. The corresponding odds ratio is $e^{.75} = 2.12$ or equally $1.078^{10} = 2.12$. We see that the odds of poor health relative to excellent health just over double for an increase of 10 years in age. For the income question we are looking at a change in X_2 of 5 units since income is measured in thousands of dollars. The income coefficient for fair health is $b_{2,2} = -.033$ so a $5000 increase in income is associated with a $5b_{2,2} = 5(-.033) = -.165$ change or a .165 decrease in the log odds of fair health relative to excellent health. Converting this to an odds ratio we have OR $= e^{-.165} = .85$ soo a $5000 increase in income is associated with a decrease of 15% in the odds of fair (vs excellent) health, all else equal.

(c) Looking at the p-values for the individual coefficients at the different levels of the model we see that age is significant for differentiating all of good health, fair health and poor health from excellent health (p-values 0, .001 and 0 respectively) after adjusting for income and depression status. Income is just barely not significant for differentiating good health from excellent health (p-value .051), just barely significant for differentiating fair health from excellent health (p-value .043) and not even close to significant for differentiating poor health from excellent health. Overall income does not seem to be nearly as important as age after adjusting for the other variables. This could be due to multicollinearity between income and age or depression. It could also be because even people with a lot of money will eventually become ill so the effect is simply not as strong. For instance the apparent anomaly with poor health could be that really poor health is a function of old age/physical deterioration and not the sort of minor effect that can be prevented by good diet, regular medical care and so on that comes with higher ses. One can ALWAYS make up a story to fit what model coefficients are telling you though–you shouldn't take any of these explanations too seriously without a lot more data or a biological mechanism! Finally, depression status does not significantly differentiate between good and excellent health (p-value .468) but does significantly differentiate between fair and excellent health (p-value = .026) and poor and excellent health (p-value = .008). These tests are all Wald tests so may be a tad on the conservative side.

To test whether age, income and depression status are significantly associated with health status across the categories we can either use likelihood ratio tests, fitting the model with and without each variable in turn, or we can use a Wald test after fitting the overall model. I demonstrate the hypotheses for the age variable. The other two are completely parallel:

$H_0 : \beta_{1,1} = \beta_{2,1} = \beta_{3,1} = 0$–age is not associated with health status after adjusting for income and depres-

sion status; age does not differentiate any of the other health levels from excellent health; age is not worth including in the model
H_A : At least one of $\beta_{1,1}, \beta_{2,1}, \beta_{3,1} \neq 0$–age is associated with health status even after adjusting for income and depression status; age separates at least one of the other health status levels from excellent health; age is worth including in the model

The printouts for both the likelihood ratio chi-squared tests and the Wald tests are shown below. We see that according to the more conservative Wald tests age and depression status are overall significant (p-values 0 and .019 respectively) while income is overall not significant (p-value .08) though it is not that far off. Using the likleihood ratio tests we get essentially the same results. Overall this confirms the impression from the individual coefficients which is that age and depression status are useful predictors but the additional explanatory power of income is doubtful.

```
WALD TESTS:
mlogit health age income depressed

. test age

 ( 1)  [1]age = 0
 ( 2)  [2]age = 0
 ( 3)  [3]age = 0

           chi2(  3) =   24.91
         Prob > chi2 =    0.0000

. test income

 ( 1)  [1]income = 0
 ( 2)  [2]income = 0
 ( 3)  [3]income = 0

           chi2(  3) =    6.79
         Prob > chi2 =    0.0789

. test depressed

 ( 1)  [1]depressed = 0
 ( 2)  [2]depressed = 0
 ( 3)  [3]depressed = 0

           chi2(  3) =    9.94
         Prob > chi2 =    0.0191
**********************************************************************
LIKELIHOOD RATIO TESTS:

.mlogit health age income depressed
. estimates store healthfull

.mlogit health income depressed
. estimates store healthnoage
```

```
.mlogit health age depressed
. estimates store healthnoincome

.mlogit health age income
. estimates store healthnodep

. lrtest healthnoage healthfull

Likelihood-ratio test                                     LR chi2(3)  =     28.93
(Assumption: healthnoage nested in healthfull)            Prob > chi2 =    0.0000

. lrtest healthnoincome healthfull

Likelihood-ratio test                                     LR chi2(3)  =      7.31
(Assumption: healthnoincome nested in healthfull)         Prob > chi2 =    0.0627

. lrtest healthnodep healthfull

Likelihood-ratio test                                     LR chi2(3)  =      9.51
(Assumption: healthnodep nested in healthfull)            Prob > chi2 =    0.0233
```

(d) Now we need to check for age and depression status whether the coefficients are the same across different levels of the model. Using depression status as an example, our hypotheses would be

$H_0 : \beta_{1,3} = \beta_{2,3} = \beta_{3,3}$–the difference in odds between depressed and non-depressed people is the same when comparing each of good, fair and poor health to excellent health
H_A : Not all of the $\beta_{j,3}$'s are the same the effect of depression on the odds of good, fair and poor health are not all the same.

Here it is easiest to use a Wald test. The corresponding printouts are shown below. I included income for reference even though it was not significant overall. Note that we have three different coefficients each for depression/age/income but we are asking if they are all the same–which would leave us with a single common coefficient. Thus this is a two degree of freedom test. We have evidence for differential effects of age and depression on the different health outcomes with p-values around .035. However there is no evidence that the effect of income differs across the levels of the model. This is hardly surprising since we didn't have any evidence that income was significant predictor (i.e. that the coefficients differed significantly from 0–and if they were all 0 they would all be equal.)

```
 test [1=2=3]: age

 ( 1)  [1]age - [2]age = 0
 ( 2)  [1]age - [3]age = 0

           chi2(  2) =    6.67
         Prob > chi2 =    0.0356

. test [1 = 2 = 3]: income

 ( 1)  [1]income - [2]income = 0
```

```
 ( 2)  [1]income - [3]income = 0

           chi2(  2) =    2.12
         Prob > chi2 =    0.3473

. test [1 = 2 = 3]: depressed

 ( 1)  [1]depressed - [2]depressed = 0
 ( 2)  [1]depressed - [3]depressed = 0

           chi2(  2) =    6.73
         Prob > chi2 =    0.0345
```

(e) From the morass of printouts above it seems that higher age is associated with higher likelihood of poorer health and that the effect gets larger the greater the health differential you are assessing (good vs excellent health, fair vs excellent health, poor vs excellent health). Thus we have a natural ordering for the age variable and the respective health levels. Similarly, being depressed is associated with higher odds of poorer health and the effect is again greater the bigger the health status differential you are considering. On the other hand, it appears that income may not have any association with health status after adjusting for age and depression status.

(f) To get a predicted probabilities in a multinomial logit you take the exponentiated log odds for that level and divide it by the sum of the exponentiated log odds for all the levels. For the reference level the exponentiated log odds is 1. For a four level model we have

$$P(Y=0) = \frac{1}{1+e^{X\beta_1}+e^{X\beta_2}+e^{X\beta_3}}$$

$$P(Y=1) = \frac{e^{X\beta_1}}{1+e^{X\beta_1}+e^{X\beta_2}+e^{X\beta_3}}$$

$$P(Y=2) = \frac{e^{X\beta_2}}{1+e^{X\beta_1}+e^{X\beta_2}+e^{X\beta_3}}$$

$$P(Y=3) = \frac{e^{X\beta_3}}{1+e^{X\beta_1}+e^{X\beta_2}+e^{X\beta_3}}$$

In this very dubious notation $X\beta_j$ is shorthand for the linear combination $\beta_{j,0}+\beta_{j,1}X_1+\beta_{j,2}X_2+\beta_{j,3}X_3$. It is easiest to calculate these systematic parts first and then plug in to the formulas above. We are given $X_1 = age = 50$, $X_2 = income = 50$ and $X_3 = depressed = 1$ since the person is 50 years old, makes \$50,000 and is depressed. The systematic pieces are

$$X\beta_1 = -.992 + .028(50) - .0177(50) + .277(1) = -.192$$

$$X\beta_2 = -2.568 + .038(50) - .034(50) + 1.083(1) = -1.282$$

$$X\beta_3 = -6.368 + /075(50) + .0027(50) + 1.794 = -.67$$

The correpsonding probabilities are

$$P(Y=0) = \frac{1}{1+e^{-.192}+e^{-1.282}+e^{-.67}} = .38$$

$$P(Y=1) = \frac{e^{-.192}}{1+e^{-.192}+e^{-1.282}+e^{-.67}} = .32$$

$$P(Y=2) = \frac{e^{-1.282}}{1+e^{-.192}+e^{-1.282}+e^{-.67}} = .11$$

$$P(Y=1) = \frac{e^{-.67}}{1+e^{-.192}+e^{-1.282}+e^{-.67}} = .20$$

The fact that these don't quite add to 1 is due to rounding. Such a person has, according to our model a 38% chance of being in excellent health, a 32% chance of being in good health, a 11% chance of fair health and a 20% chance of poor health. The fact that the chance of poor health is higher than that of fair health is mostly due to the fact that the pattern in the income variable is inconsistent and this person has a fairly high income.

(g) The ordinal logistic printout is shown below. To see whether it is overall significant we look at the likleihood ratio chi-squared test which has a p-value of 0 so at least one of age, income or depression status is related to health status according to this model.

```
. ologit health age income depressed

Ordered logistic regression                     Number of obs   =        294
                                                LR chi2(3)      =      40.81
                                                Prob > chi2     =     0.0000
Log likelihood = -310.73771                     Pseudo R2       =     0.0616

------------------------------------------------------------------------------
      health |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0312823   .0065539     4.77   0.000      .018437    .0441277
      income |  -.0175341   .0082107    -2.14   0.033    -.0336269   -.0014413
   depressed |   .8128966   .3088315     2.63   0.008     .2075981    1.418195
-------------+----------------------------------------------------------------
       /cut1 |   .8924209   .3918619                      .1243857    1.660456
       /cut2 |   2.939776   .4307967                      2.095429    3.784122
       /cut3 |    4.39194   .4938195                      3.424071    5.359808
------------------------------------------------------------------------------
```

(h) In the standard formulation of an ordinal logistic regression the coefficients and odds ratios for the X variables tell you how the predictors are associated with the odds of **lower** vs **higher** values of Y and the intercepts calibrate the respective probabilities of the different categories. However STATA reverses the signs on its coefficients so that a positive value of a coefficient or an odds ratio higher than 1 corresponds to a higher probability of a higher value of Y. Here we are thinking of health status being ordered from best to worst with $Y = 0$ corresponding to excellent health and $Y = 3$ correpsonding to poor health. Thus a positive coefficient in STATA for this model would tell us that **increasing** values of X are associated with higher probabilities of **poorer** health. Thus the coefficient of age, $b_1 = .031$ tells us that each extra year of age is associated with an increase of .031 in the log odds of poorer health, all else equal. Under the proportional odds assumption this change in log odds is the same whether we are comparing not excellent health to excellent health, fair or poor health to good or excellent health, or poor health to not poor health. The corresponding odds ratio is $e^{.031} = 1.03$ which means the odds of poorer health go up 3% per extra year of age. Age is a highly significant predictor with a p-value of 0.

The coefficient of the income variable tells us that the log odds of poorer health go down .0175 points per extra $1000 of income. The corresponding odds ratio is $e^{-.0175} = .983$. In other words there is a 1.7% reduction in odds of poorer health for each extra $1000 in income, all else equal. The p-value for income is .033 which is significant though not as strongly for age.

Finally, the coefficient of the depression indicator $b_3 = .813$ means that the log odds of poor health are .813 higher for depressed people than non-depressed people of the same age and income. The correponding odds ratio is $e^{.813} = 2.25$ which means the odds of poorer health are over twice as high for depressed people than non-depressed people, all else equal. The depression variable is highly significant with a p-value of .008.

(i) Now we need to get the predicted probabilities for the ordinal logistic regression. The way the ordinal logistic model is set up we have

$$ln(\frac{P(Y=0)}{P(Y\neq 0)}) = \beta_{0,0} + \beta_1 X_1 + \cdots$$

In analogy with logistic regression it follows that

$$p_0 = P(Y=0) = \frac{e^{X\beta_0}}{1+e^{X\beta_0}}$$

where by $X\beta_0$ I mean the expression on the right side of the first equation above. Similarly we have

$$ln(\frac{P(Y=0,1)}{P(Y\neq 0,1)}) = \beta_{1,0} + \beta_1 X_1 + \cdots$$

To get $p_1 = P(Y=1)$ we just need to subtract the probability that $Y=0$ from the probability that $Y=0$ or 1. We continue this way sequentially. The final thing we need to keep in mind is how our package reports the assorted coefficients. As I noted in part (h) STATA's slope coefficients correspond to the log odds of **higher** values of Y. Thus for the probability calculations above which focus on getting the lower values of Y we need to take negatives. STATA's cut points correspond to the constant terms. Thus for predicting the probability of excellent health for a 50 year old man with a $50,000 income who is depressed we get the following linear combination of the X values, remembering to reverse the signs on our coefficients. Our value for $\hat{\beta}_{0,0}$ is what STATA calls cut1:

$$.892 - .031(50) + .0175(50) - .813(1) = -.621$$

The corresponding predicted probability is

$$\frac{e^{-.621}}{1+e^{-.621}} = .350$$

In a similar manner we get the linear combination for the probability that $Y=0$ or $Y=1$ using cut 2 as

$$2.94 - .031(50) + .0175(50) - .813(1) = 1.427$$

Note that this isn't as big a hassle as with the multinomial model since the coefficients are the same each time–all that changes is the constant term! The correpsonding probability is

$$\frac{e^{1.427}}{1+e^{1.427}} = .806$$

It follows that $p_1 = .806 - .350 = .456$.

Proceeding with cut 3 we get the linear combination for the probability of $Y=0$ or $Y=1$ or $Y=2$ as

$$4.39 - .031(50) + .0175(50) - .813(1) = 2.877$$

With corresponding probability

$$\frac{e^{2.877}}{1+e^{2.877}} = .947$$

Again, by subtraction we get $P(Y = 2) = .947 - .806 = .141$. Finally, $P(Y = 3) = 1 - .947 = .053$. Thus a depressed 50 year old with an income of \$50,000 has a 35% chance of excellent health, a 45.6% chance of good health, a 14.1% chance of fair health and a 5.3% chance of poor health.

(j) In the ordinal logistic model all three of our predictors are significant an in the expected direction so conceptually this is a little more appealing than the multinomial model in which the income variable behaved a bit strangely. The probabilities of the respective categories also seem a little more natural. To formally compare the multinomial and ordinal models we look at the log likelihoods. The log likelihood for the multinomial model is -307.02012 while that for the ordinal logistic model is -310.73771. Note that the multinomial model HAS to have a better log likelihood because it is more general–the question is whether it is enough better to justify all the extra parameters. In this case with 3 predictors and 4 levels we have 6 extra parameters in the multinomial model (both models have a different intercept for each level but the multinomial model has two extra slope terms for each variable.) Although the models are not exactly nested -2 times the difference in log likelihoods is still fairly close to having a chi-squared distribution with degrees of freedom equal to the difference in number of model parameters. Here this difference would be 7.44. We can use the chi2tail command from homework 2 to get the associated p-value as follows:

```
. display chi2tail(6, 7.44)
.282064
```

This p-value is fairly large so it doesn't look like the more flexible multinomial model is a significant improvement over the simpler ordinal logistic model and we can safely use the results from the parts (g)-(i) which are more intuitive.

(k) The proportional odds statement says that the coefficients giving the effect of a predictor on the log odds of lower (vs higher) values of the outcome variable are the same, regardless of where one draws the line between lower and higher values (and hence the odds ratios for the predictors are also the same across all splits of the outcome.) In this problem, our outcome variable is health status, which has four levels ranging from 0 = Excellent Health to 3 = Poor Health, and our predictors are age, income and depression status. The proportional odds assumption in this context means that the effects of being older, having a higher income and being depressed on the odds of better (vs worse) health are the same, regardless of whether one defines "better health" as Excellent only; Excellent or Good; or Excellent, Good or Fair. There are a number of tests of the proportional odds assumption. The first piece of STATA output below gives 5 such tests, of which the third, the score test, is the one given by default in SAS. All of these tests are very non-significant, meaning we have no evidence of a significant violation of the proportional odds assumption. This is consistent with our earlier results suggesting that the ordinal model is probably adequate for these data, i.e. that the multinomial model did not provide a big improvement. The second piece of STATA output, which provides details of the Brant test, gives the estimated coefficients (and test statistics) one would get if one actually ran logistic models with the splits as specified by the different cutpoints. From this we can see that in most cases the estimates are very similar, as they should be if the proportional odds assumption is correct, although the depression variable coefficient seems to perhaps be slightly different at the first cutpoint than at the other two. The second part of the output gives both an overall test and tests for the individual variables. Again, none of them are significant, reinforcing that the small differences we see in the coefficients are not enough to cast serious doubt on the proportional odds assumption.

```
. oparallel

Tests of the parallel regression assumption

                 |   Chi2     df  P>Chi2
-----------------+----------------------
     Wolfe Gould |  1.392      6   0.966
           Brant |  5.892      6   0.435
           score |  5.452      6   0.487
likelihood ratio |  5.517      6   0.479
            Wald |  5.452      6   0.487
****************************************
. brant, detail

Estimated coefficients from binary logits

----------------------------------------------
    Variable |  y_gt_0     y_gt_1     y_gt_2
-------------+--------------------------------
         age |    0.033      0.032      0.055
             |     4.43       3.38       3.06
      income |   -0.019     -0.014      0.015
             |    -2.26      -1.09       0.71
   depressed |    0.586      1.142      1.456
             |     1.69       2.95       2.33
       _cons |   -0.915     -3.103     -6.469
             |    -2.19      -4.93      -4.87
----------------------------------------------
                                  legend: b/t

Brant test of parallel regression assumption

             |       chi2     p>chi2      df
-------------+------------------------------
         All |       5.89      0.435       6
-------------+------------------------------
         age |       2.01      0.365       2
      income |       3.14      0.208       2
   depressed |       2.36      0.307       2

A significant test statistic provides evidence that the parallel
regression assumption has been violated.
```