Solutions To Homework Assignment 3 Warm-Up Problems

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).