Solutions To Homework Assignment 2

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) Don't Drink and Derive: The purpose of this problem is to see some examples of testing and interpreting transformations and interactions in the logistic regression context.

(a) Interactions We start by interpreting the interaction effect of speed and whether or not it was dark at the time of the accident on whether or not there were fatalities in the accident. The intuitive definition of an interaction is that the relationship between the outcome and the first predictor variable **depends on or varies as a function of** the value of the second predictor variable and vice versa. Here what that means is that the relationship between the likelihood of fatalities and speed "depends on" whether or not it was dark out (i.e. is different for day and night driving) or equivalently the difference in likelihood of fatalities between night and day changes as a function of how fast the cars were going. The first of these versions of the interpretation is probably more intuitive but both are perfectly valid and both make common sense. Driving fast may be more dangerous at night (version 1) or the effects of night driving (reduced vision, etc.) may be exaccerbated by higher speed (version 2). In ordinary least squares regression, if you have one indicator (group) variable and one continuous variable and no interaction then essentially you are fitting two parallel regression lines, one for each group. They may have different intercepts but the slopes are the same. If you include an interaction term then you are allowing the lines to have different slopes as well as different intercepts. This can be extended to the case where either both predictor variables are categorical or both are continuous but the one categorical/one continuous case is the easiest to visualize. Mathematically, to include an interaction term in the model we take the product of the two variables involved in the joint effect. The coefficient of the interaction term gives the difference in slope between the two lines. So how does this generalize to the logistic regression setting? In logistic regression everything is linear on the **log odds scale**. Thus we can interpret interactions in logistic regression the same way as in OLS as long as we talk about the outcome in terms of log odds. If we go to the odds ratio scale then we can still say that the odds ratio for the effect of speed is different for nighttime driving than daytime driving or that the odds ratio for driving at night changes as a function of speed. However in this case the exponentiated coefficient of the interaction term has a **multiplicative** effect on the odds ratio.

The most important thing to remember with interaction terms, in logistic regerssion as well as OLS, is that you can't interpret the coefficients (or odds ratios) for the main effects of the componment variables separately–you must interpret them in terms of the combined effects of the two (or more) variables involved in the interaction. All variables not involved in the interaction are assumed to be held fixed and so don't really affect the interpretation of the interaction.

As an illustration, let's look at the coefficients for the speed and dark variables and their interaction in the traffic data set. I'm going to assume that the at fault driver was male and not drunk and that there was a two-car accident so that I can set $X_1 = X_2 = X_8 = X_9 = 0$. I will also assume the driver was 40 years old. (A standard choice is to set all the continuous variables not involved in the interaction to their mean value in the data set. I just picked a round number near that.) None of these values actually matters for the interpretation of the interaction–they just change what is my base value for the log odds (the intercept for the plot I'm going to draw). We could equally carry the variables along in our expressions without plugging in values but it's messier looking. Now let's think about the values $b_5 = .026$, the speed coefficient, $b_6 = 2.73$, the dark coefficient, and $b_7 = .018$, the interaction coefficient. At an arbitrary speed and time of day, using the values specified above for the other variables, my equation for the log odds is

$$-3.367 + .896(0) - .967(0) - .15(40) + .0015(40^2) + .026speed + 2.73dark + .018speed * dark + 1.18(0) + .95(0)$$

$$= -6.97 + .026speed + 2.73dark + .018speed * dark$$

The simplest way to interpret the remaining coefficients is to look at what happens (1) during the day (dark = 0) at an arbitrary speed and (2) at night (dark = 1) at an arbitrary speed. The equation for the log odds as a function of speed during the day is just

$$-6.97 + .026speed + 2.73(0) + .018speed(0) = -6.97 + .026speed$$

In other words our "intercept" of -6.97 gives the log odds for fatality during the day time if our speed is 0 (i.e. basically stopped) and the coefficient b_5 for the speed variable gives the change in log odds associated with driving 1 extra mile per hour faster **during the day**. Similarly the odds ratio obtained by exponentiating the speed coefficient, $e^{b_5} = 1.026$ says the odds of a fatality are 2.6% higher per mile per hour of speed **assuming we are driving during the day**.

Now what about at night time? In this case our equation becomes

$$-6.97 + .026speed + 2.73(1) + .018speed(1) = (-6.97 + 2.73) + (.026 + .018)speed = -4.24 + .044speed$$

We see that the coefficients of the dark indicator and the interaction term give us, respectively, the differences in intercept and slope between driving at night and driving during the day. Specifically, $b_6 = 2.73A$ tells us that if the accident takes place at 0 miles per hour the log odds of a fatality will be be 2.73 higher if the accident happens at night compared to during the day. On the odds ratio scale, $e^{2.73} = 15.27$ which means that for accidents that happen at 0 speed the odds of a fatality are 15 times as high at night as during the day. This is not really very meaningful since you don't have accidents at 0 miles per hour and even if you take this to mean extremely low speed the likelihood of a fatality is so low either during the day or at night that we probably aren't going to be considering it. The intercepts here are more important for giving the baseline on which to build the effects of the other variables. Now let's consider the coefficient of speed in the night-time equation. It is $.026 + .018 = .044$ meaning that at night every extra mile per hour of speed increases the log odds of a fatality by .044. The coefficient of the interaction term, $b_7 = .018$ tells us how much more the log odds go up by at night than during the day. If we exponentiate to get the odds ratio we have $e^{.044} = 1.045$ meaning that **at night** each extra mile per hour is associated with a 4.5% increase in the odds of a fatality. Note that $e^{.045} = e^{.026+.018} = e^{.026}e^{.018} = (1.0265)(1.0185)$ which is the product of the odds ratios given on the printout for the speed main effect and the interaction term. In other words, the odds ratio for the interaction term tells us the extra multiplicative effect on the odds of fatality associated with driving at night as opposed to during the day. The effects of all these components are even messier on the probability scale. Because the probabilities are non-linear, the change in probability depends not only on the coefficients/odds ratios but the actual values of the speed and dark variables. The easiest way to see the effects is to plot the predicted probabilities as a function of speed for (1) day time and (2) night-time

and to see where the resulting S-curves differ.

(b) Transformations: Transformations of the predictor variables in logistic regression work the same way as in standard linear regression: you add functions of the X's as predictors, test whether they are significant, and if so interpret them in terms of the **shape** of the relationship between the predictor and outcome (rather than in terms of the effect of a 1-unit change in X since for non-linear transformations the degree of impact depends on the actual X value.) The main difference is that the interpretation about the shape is simplest on the log odds scale (where it is the same as in OLS) but that is the least intuitive scale in terms of the outcome. The numeric interpretation on the odds ratio scale is NOT so easy but computer packages will provide odds ratios for the coefficients of transformed variables anyway–you don't want to interpret them blindly. Usually you can at least say something about the effect on the probability. For instance if you have a quadratic predictor that is significant with a positive coefficient that means that the probability of the event of interest initially goes down as X increases and then goes up again–it's just that the shape isn't a parabola on the probability scale. If you want odds ratios associated with a curvilinear variable it's best to pick two specific values of X you'd like to compare, compute the difference in the log odds at those two values and exponentiate that to get the odds ratio of interest.

As a specific example let's look at age as a predictor in the traffic accident data set. The model includes a linear and a quadratic term and the coefficient of the age-squared term is positive. This means that we expect the log odds (and hence the probability) of a fatality to be higher for both younger and older drivers and lower for drivers of middle age. This makes intuitive sense–inexperience and a higher tendency towards recklessness on the young end and deterioration of reaction times and other phyical faculties on the high end could both lead to worse accidents. To formally test whether the quadratic relationship between age and log odds of fatality is an improvement over a linear relationship we just need to test whether the coefficient of the age-squared term is significantly different from 0:

$H_0 : \beta_4 = 0$–the age-squared term contributes nothing to the model after accounting for the other variables; a quadratic relationship is no better than a linear relationship in age.
$H_A : \beta_4 \neq 0$–the age-squared term is worth adding to the model; a quadratic effect better describes the relationship between age and log odds of fatality than a linear relationship.

The p-value for the corresponding Wald test is .005 so we reject the null hypothesis and conclude that the quadratic model is an improvement. However, how much the log odds or the probability changes with age changes depends on where you are along the parabola. Suppose for instance that we want to compare the odds of a fatal accident for drivers at two different ages, A_1 and A_2. The difference in log odds for these two people is given by

$$b_4A_1 + b_5(A_1^2) - b_4A_2 - b_5(A_2^2) = -.15(A_1 - A_2) + .0015(A_1^2 - A_2^2)$$

If we compare a 30 year old to a 20 year old our change in log odds is

$$-.15(30 - 20) + .0015(900 - 400) = -.75$$

and the corresponding odds ratio is $e^{-.75} = .47$. The 30 year old has about 53% lower odds or odds only about half as high odds of being in a fatal accident as the 20 year old, all else equal. However if we compare a 40 year old to a 30 year old, which is also a 10 year change in age, the difference in log odds is

$$-.15(40 - 30) + .0015(1600 - 900) = -.45$$

and the corresponding odds ratio is .64. Thus a 40 year old has only a 36% reduction in odds of a fatality compared to a 30 year old. There is no easy direct way to interpret the odds ratio of .0015 given on the printout for age-squared because it can not be separated from age. Note that this is different from the interaction situation above where we were able to get separate interpretations by breaking the equation down

into cases (day vs night). Here we have to give the interpretation for a particular pair of ages.

We could conceptually have come up with the idea that a quadratic model in age would make sense here but suppose we weren't sure about the shape and wanted to find a way to visualize the data, similar to a scatterplot of Y vs X in OLS. One way this is done is to divide the range of the X variable (here age) into intervals (quartiles is a common choice), create dummy variables for each of the intervals, fit a logistic regression using those dummy variables, and then to plot their corresponding coefficients versus the middle of the bins. You plot a coefficient of 0 for whichever bin you used as the reference. Since the coefficients are on the log odds scale, if this plot looks linear then you suspect a linear model will be OK. If the plot looks curved it provides some intuition about the shape. Of course the more bins you use the better the idea you get of the shape but also the more data you need to get a reasonable fit with all those dummy variables!

(2) Goodness of Fit Basics:

(a) In a logistic regression we model the **probability** of an event of interest as a function of a set of predictor variables. The probability represents the **average** relationship between the outcome and the covariates for the population as a whole. You can use the predicted probabilities together with a threshold cutoff to make a prediction about the outcome for an **individual** subject. The higher the predicted probability, the more inclined you are to believe that the individual will have the event. The term **calibration** refers to how well your model fits **on average**–in other words it measures how good your fit is on the level of the **probabilities**. Suppose we had as our predictors in a logistic model two indicators, say gender and treatment group. Our model would produce four predicted probabilities–one for male and treated, one for male and untreated, one for female and treated and one for female and untreated. We could also compute the observed proportions of events in each of these four categories in our data set and see if they were similar to the predicted probabilities. The closer they are, the better our model fits. This is the essential idea behind the Hosmer-Lemeshow goodness of fit test. HL simply extends this idea to the case where we have continuous predictors so it is not as obvious what our bins of covariate values should be. **Predictive accuracy** refers instead to how good a job our model does at guessing the outcomes for **individuals**. This is obviously a much harder problem. Even if we have accurately estimated the probability for people with a certain set of characteristics as p, the chances that a particular individual with those characteristics has the event will be p and the chance that they won't will be 1-p. Wherever we set our threshold we will make a mistake on a fraction of either p or 1-p of the people. The closer p is to .5 the more errors we will tend to make because about half the people will have the event and half will not but we will categorize them all as having it (or not having it) depending on where we've set our threshold. This is somewhat analagous to the situation in standard linear regression where even if we had estimated the regression line perfectly, individuals would still vary about the line, e.g. it doesn't have to be the case that all people with the same height (X) have the same weight (Y) even if height and weight are related. However in OLS the values for the average Y and the individual Y (here weight) are at least on the same scale. In the logistic setting the average outcome is a probability whereas the individual outcome is a yes or no. It is possible to have good calibration and bad predictive accuracy (e.g. you fit the right model but lots of people in your sample have characteristics that make their predicted probabilities near .5). It is also possible to have bad calibration (i.e. your predicted probabilities are way off) but good predictive accuracy because the people who do and do not have the outcome of interest are well-separated in terms of their X characterstics and hence easy to distinguish from one-another. Thus these two concepts tell you different things about how well your model performs.

(b) The terms **sensitivity** and **specificity** are connected to the concept of predictive accuracy. **Sensitivity** is also called the **true positive rate**. It is the fraction of people who have the event of interest who you correctly predict to have the event using your model. The **specificity** is also called the **true negative rate**. it is the fraction of people who do NOT have the event who you predict correctly with your model. Instead of specificity people sometimes talk about the **false positive rate** which is just 1 minus the specificity. The sensitivity and specificity depend on the probability cutoff or **threshold** you use for deciding whether to

predict the person will have the event. For a threshold, p_0, you predict the person will have the event if their predicted probability is above the cutoff, i.e. $\hat{p} > p_0$ while if $\hat{p} < p_0$ you will predict the person doesn't have the event. The smaller you set your p_0, the higher your sensitivity will be. In fact, if $p_0 = 0$ you will predict that everyone will have the event and hense your sensitivity will be 1–all people who actually have the event will have been predicted to do so. However this naturally does horrible things to your specificity. If you predict that everyone will have the event then you will be wrong about 100% of people who don't have the event and your specificity will be 0. You can of course make your specificty higher by increasing the threshold probability but at the expense of decreased sensitivity. The trick is to find a point at which sensitivity and specificity are both as high as possible. This is often done by plotting both sensitivity and specificity as a function of the probability threshold and picking the point where the curves cross. However it may be more important to you to classify one or the other of the outcomes correctly (e.g. it might be worse to miss diagnosing someone who has a disease than to say someone has the disease when they don't. In the first instance the person may go off an infect other people or not get the needed treatment while in the second further tests will presumably show the person is OK though there may be some un-necessary procedures/anxiety in the interim). In this case you might choose to optimize one of sensitivity or specificty over the other. The combined sensitivity and specificity curves can be summarized using something called a **receiver operating characteristic** or **ROC** curve. This curve plots sensitivity (y-axis) versus false positive rate (x-axis) over a range of threshold probabilities. If your model is no better than random guessing the ROC curve will look like a straight line with a slope of 1 going from (0,0) to (1,1). An optimal ROC curve will rise rapidly to 1 and level off–this corresponds to being able to achieve high sensitivity even while mainintaining a low false positive rate (high specificity). The information from the ROC curve is usually summarized by the **area under the curve** or **AUC** which is 1 for a perfect model and .5 for a model that is no better than random guessing. Values in the .7 to .9 range are generally considered pretty good. Achieving values above .9 is very rare.

(3) Still Gasping For Breath: This problem is designed to illustrate the various goodness of fit measures, including examining the shape of the relationship for a continuous variable, calibration and predictive accuracy.

(a) To check for linearity of a predictor in a logistic regression a rough but fairly simple technique is to divide the variable into quartiles, create dummy variables for the resulting bins and run a logistic model using those dummy variables. One then plots the resulting coefficients for the indicators versus the mid-point of the bins. If the plot looks linear then a linear fit is probably OK. If it has a curved pattern than will indicate something about a superior shape. The commands used to obtain the quantiles (25th, 50th and 75th percentiles) for the pollution variable are shown below. The resulting bins are [0-15.7], [15.7-21.3], [21.3-26.1] and [26.1,40.5]. We then fit the model with all the predictors including the dummy variables for these ranges. The resulting coefficients are $\hat{\beta}_1 = .51, \hat{\beta}_2 = .85, \hat{\beta}_3 = 1.25$. The lowest pollution bin is serving as the reference so effectively its coefficient is 0. We use this value in the scatterplot which is included in the accompanying graphics file. The points in the plot fit a straight line very well so we conclude that a linear fit in pollution is appropriate. The printout for the logistic regression with a linear term in pollution is also given below. We will use this model as the basis for the rest of the problem.

```
. sum pollution, detail

                        pollution
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4.581451       .6100724
 5%     9.353241       .6554115
10%     11.36944       .9150193       Obs                1000
25%     15.69217       1.728456       Sum of Wgt.        1000
```

```
50%     21.25496                    Mean           21.00246
                      Largest       Std. Dev.      7.221641
75%     26.08167      39.60468
90%     30.31795      39.82299      Variance        52.1521
95%     32.43209       39.8604      Skewness       -.0813572
99%     37.01562      40.41085      Kurtosis        2.622869
*****************************************************************
To create the pollution bin indicators I proceeded as follows:

gen pollbin1 = 0
replace pollbin1 = 1 if pollution > 15.7 & pollution <=21.3

and so on for the other indicators

. logit asthma pollbin1 pollbin2 pollbin3 urban ses breastfed famhist genderp3

Logistic regression                               Number of obs   =       1000
                                                  LR chi2(8)      =      76.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -334.50151                       Pseudo R2       =     0.1029

------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    pollbin1 |   .5072169   .3756775     1.35   0.177    -.2290975    1.243531
    pollbin2 |   .8546278   .4145998     2.06   0.039     .0420272    1.667228
    pollbin3 |   1.254061   .4366646     2.87   0.004     .3982144    2.109908
       urban |    .045625   .3151697     0.14   0.885    -.5720963    .6633462
         ses |  -.0821695   .0347055    -2.37   0.018    -.1501911   -.0141479
   breastfed |   .0042069    .010241     0.41   0.681     -.015865    .0242788
     famhist |   1.171824   .2225175     5.27   0.000     .7356974     1.60795
    genderp3 |  -1.080269   .2166788    -4.99   0.000    -1.504952   -.6555864
       _cons |   .9074093   1.622919     0.56   0.576    -2.273454    4.088272
------------------------------------------------------------------------------
*******************************************************************************

To create the variables coeff and midpoint I simply opened the data editor
and created two new columns by typing in the numbers--specifically the
estimated betas (with 0 in the place for pollbin0 since it was the
reference) and the midpoints of my bins (e.g. 7.85 as the midpoint of the
bin 0-15.7 and so on).  Then I created the scatterplot:


. scatter coeff midpoint

. logit asthma urban pollution ses breastfed famhist 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
     famhist |   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
------------------------------------------------------------------------------

. logistic asthma urban pollution ses breastfed famhist 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 | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   .9257104   .2796262    -0.26   0.798     .5121009     1.67338
   pollution |    1.07965   .0216071     3.83   0.000      1.03812     1.12284
         ses |    .921448   .0318649    -2.37   0.018     .8610635    .9860672
   breastfed |   1.005744   .0103814     0.55   0.579     .9856018    1.026299
     famhist |    3.25478   .7259464     5.29   0.000     2.102178    5.039341
    genderp3 |   .3346326     .07291    -5.02   0.000     .2183271    .5128953
------------------------------------------------------------------------------
```

(b) The Hosmer-Lemeshow test measures goodness of fit by dividing the subjects into bins, counting how many subjects in each bin have or do not have the event of interest (the "observed counts"); computing the probabilities of having or not having the events for the subjects in each bin based on the fitted model; summing them (to get the "expected counts"); and comparing the resulting values using a statistic analagous to the contingency table chi-squared statistic. If the observed counts are very similar to the expected counts calculated from the model then the model fits well (specifically is "well-calibrated"). If the observed counts are significantly different from the expected counts it means there are at least some bins in which the model does not fit well. If the predictor variables are all categorical then one can let the resulting set of possible predictor categories represent the bins. However if some of the predictors are continuous one usually divides the subjects into bins based on quantiles of the predicted probabilities. 10 bins is a standard choice. The printout for the H-L test for the asthma model is given below. The p-value for the chi-squared test is .11 which is not significant. This suggests that our model fits reasonably well–the observed values do not differ significantly from what is suggested by our model.

```
. estat gof, group(10)

Logistic model for asthma, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)
```

```
 number of observations =      1000
       number of groups =        10
Hosmer-Lemeshow chi2(8) =        12.93
            Prob > chi2 =         0.1144
```

(c) The pseudo-R^2 based on the log likelihood is given in the basic logistic printout. It is only .11 for this model, indicating that despite the enormous overall significance of the model and the individual predictors, and the fact that the model fits quite well (per the HL test), we are still a long way from a perfect fit. If you like to think of it this way we have moved only 11% of the way along the path from the log-likelihood of the null model (no predictors) to the saturated model (perfect fit.) I mentioned other pseudo-R^2 measures in class. One is the square of the correlation between the Y values and the predicted probabilities. The commands for obtaining this value are shown below. The resulting pseudo-R^2 is $r^2 = .27^2 = .073$ which is a little worse than the .11 we get from the likelihood approach. Another option is to use sums of squares, specifically

$$1 - \frac{\sum(y_i - \hat{p}_i)^2}{\sum(y_i - \bar{y})^2} 2 = 1 - \frac{1001.45}{1078.7} = .072$$

This is about the same as the squared correlation. I got the values for the sums of squares by creating two new variables, the squared residuals based on the model and the squared residuals based on the null model (which just predicts the overall fraction of the cases in the sample as the probability for everyone). The ratio of the means for these variables is the same as the ratio of the sums of squares since the sample size is the same in each case. The corresponding output is shown below. This metric produces an even lower pseudo-R^2 than the other two measures. The main point here is to recognize that although the model fits well and is significant there is still a lot of unexplained variability–this is very typical of logistic regression. Moreover there are many different measures of R^2 and they are not particularly consistent in their numerical values. These measures are most useful for comparing different models with the same outcome variable to see how much better one performs than another.

```
Correlation pseudo R-squared

. predict asthmaprobs
(option p assumed; Pr(asthma))

. cor asthma asthmaprobs
(obs=1000)

             |   asthma asthma~s
-------------+------------------
      asthma |   1.0000
 asthmaprobs |   0.2698   1.0000
****
. gen asthmaresids = asthma - asthmaprobs

. summarize asthma

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      asthma |      1000        .123    .3286016          0          1
```

```
. gen asthmaresidsnull = asthma - .123

. gen asthmaresids2 = asthmaresids^2

. gen asthmaresidsnull2 = asthmaresidsnull^2

. summarize asthmaresids2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
asthmares~s2 |      1000    .1001454    .2208187    .000085   .9294102

. summarize asthmaresidsnull2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
asthmares~l2 |      1000     .107871    .2477656    .015129    .769129
****************************************
```

(d) The desired sensitivity, specificity and total fraction correct plots are shown in the accompanying graphics file. The first two are generated automatically by most programs. The total fraction correct can be calculated from the sensitivity and specificity as follows. Let n_P be the number of positive cases in the sample and let n_N be the number of negative cases with $n = n + P + n_N$. Let n_{TP} be the number of positive cases that are predicted correctly (the true positives) and let n_{TN} be the number of negative cases predicted correctly (the true negatives). The sensitivity is nTP/n_P and the specificity is n_{TN}/n_N. The fraction correct is

$$\frac{n_{TP} + n_{TN}}{n} = \frac{n_P(n_{TP}/n_P) + n_N(n_{TN}/n_N)}{n} = \frac{n_P}{n} Sensitivity + \frac{n_N}{n} Specificity$$

The commands for generating this variable are also shown below. The sensitivity/specificity plot shows the two curves crossing at a threshold of about .13 and the corresponding sensitivity and specificity (and hence total fraction correct) are both around 70% which is quite good. Correspondingly the area under the ROC curve is .7528 which is very good though not absolutely spectacular. Overall this model performs very well in terms of predictive accuracy, despite the rather low seeming pseudo-R^2 values.

```
. lsens, gensens(asthmasens) genspec(asthmaspec) genprob(asthmathreshold)
obs was 1000, now 1002

. tab asthma

     asthma |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        877       87.70       87.70
          1 |        123       12.30      100.00
------------+-----------------------------------
      Total |      1,000      100.00

. gen totcorrect = .123*asthmasens + .877*asthmaspec
. scatter totcorrect asthmathreshold
. lroc
```

```
Logistic model for asthma

number of observations =     1000
area under ROC curve   =   0.7528
```

Turn-In Problems

(4) Drink is a Downer:

(a) The simple logistic regresison of drink on sex is shown below. To get the odds for men and women we simply plug in the values 1 (for female) and 0 (for male) to get the log odds and then exponentiate. We have the odds for men as $e^{\hat{\beta}_0} = e^{1.78} = 5.93$ and for women the odds are $e^{\hat{\beta}_0+\hat{\beta}_1(1)} = e^{1.78-.63} = 3.16$. We can get the odds ratio for men versus women as directly as $OR = 3.16/5.93 = .53$, so women have about half the odds of being drinkers as men. Alternatively of course we could use the logistic command to get STATA to print the odds ratio or we could get it by exponentiating the sex coefficient, $e^{-.63} = .53$. The sex variable is significant (though barely with p-value .049) so it looks like there is a difference in odds of drinking between men and women with women having the lower odds.

```
. logit drink sex

Logistic regression                               Number of obs   =        294
                                                  LR chi2(1)      =       4.10
                                                  Prob > chi2     =     0.0429
Log likelihood = -146.71787                       Pseudo R2       =     0.0138

------------------------------------------------------------------------------
       drink |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.6310039   .3208548    -1.97   0.049    -1.259868     -.00214
       _cons |   1.781288   .2702338     6.59   0.000      1.25164    2.310937
------------------------------------------------------------------------------
```

(b) The separate printouts for people who are and are not depressed are shown below. Among people who are depressed the odds ratio for being a regular drinker in women vs men is 1.17 (implying women have 17% higher odds of being regular drinkers) but the odds ratio is not remotely close to significant so we have no evidence of any sex difference among depressed people. (Note however that with only 50 depressed people our power isn't that great!) Among the not depressed people the odds ratio for women vs men is .46 meaning that women have a 54% reduction in the odds of being regular drinkers relative to men. This odds ratio is significant (p-value .025 by the Wald test or .021 by the likelihood ratio chi-squared test).

```
. logistic drink sex if depressed==1

Logistic regression                               Number of obs   =         50
                                                  LR chi2(1)      =       0.03
                                                  Prob > chi2     =     0.8555
Log likelihood = -23.553082                       Pseudo R2       =     0.0007
```

```
------------------------------------------------------------------------------
       drink | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   1.178571   1.052931     0.18   0.854     .2045953    6.789162
------------------------------------------------------------------------------

. logistic drink sex if depressed==0

Logistic regression                               Number of obs   =        244
                                                  LR chi2(1)      =       5.36
                                                  Prob > chi2     =     0.0206
Log likelihood = -122.40464                       Pseudo R2       =     0.0214

------------------------------------------------------------------------------
       drink | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .4610127   .1592889    -2.24   0.025     .2342104    .9074437
------------------------------------------------------------------------------
```

(c) Overall the results of part (b) suggest that depressed men and women are equally likely to drink while not depressed women are less likely to drink than non-depressed men. This reads exactly like an interaction effect–whether there is a relationship between sex and drinking status depends on whether or not the people involved are depressed or not. In general an interaction means that the relationship between the outcome and predictor 1 depends on the value of predictor 2 and vice versa. When both variables are indicators in a logistic regression this amounts to saing that the odds ratio for predictor 1 differs for the two values of predictor 2 and vice versa. So here we would say in general that the odds ratio comparing the relative likelihood of men and women being regular drinkers is different for depressed people than for non-depressed people, or equivalently that the odds ratio comparing the relative likelihood of depressed and non-depressed people being regular drinkers is different for men and women. You can get an interaction just by having different magnitudes for the odds ratios. However here the implication is actually that for one of the groups there is no difference in odds (or, after doing the algebra, in probabilities) between men and women while in the other group there is. In fact, another way to think of an interaction of two indicators in a logistic regression is that with the interaction you can have an arbitrary odds or probability for each of the four groups (female-depressed, male-depressed, etc.) while without the interaction you are constraining the odds/probabilities (i.e. so that the odds ratios are the same across depression categories.) To tell whether this effect is real we perform a test of the sex by depression interaction in a logistic regression including both variables. The corresponding printout is shown below. Our hypotheses are

$H_0 : \beta_3 = 0$–the sex by depression interaction is not significant; the odds ratios for sex are the same whether or not the people are depressed.
$H_A : \beta_3 \neq 0$–the sex by depression interaction is significant; there is a difference in the sex odds ratios between depressed and not depressed people.
From the printout the p-value for the Wald test for the interaction is .327 which is not significant. Therefore we do not have sufficient evidence to conclude that there is an interaction and we can not be sure that the odds ratios comparing the likelihood of regular drinking in women vs men are different for people who are and are not depressed. This may seem surprising given what we saw in the first parts of this problem. However the sample size for the depressed group is really quite small–there are only 50 depressed people and most of them are drinkers so we have very little data to really work out the odds ratio for men and women in that subsample. That makes it difficult for us to have a high degree of confidence when we make the comparison to the (much larger) not depressed subsample.

```
. logit drink sex depressed depbysex

Logistic regression                                     Number of obs   =        294
                                                        LR chi2(3)      =       5.62
                                                        Prob > chi2     =     0.1318
Log likelihood = -145.95772                             Pseudo R2       =     0.0189

------------------------------------------------------------------------------
       drink |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.7743296   .3455196    -2.24   0.025    -1.451536   -.0971237
   depressed |  -.4405564   .8413815    -0.52   0.601    -2.089634    1.208521
    depbysex |   .9386327   .9578851     0.98   0.327    -.9387877    2.816053
       _cons |   1.826851   .2879632     6.34   0.000     1.262453    2.391248
------------------------------------------------------------------------------
*******************************************************************************
Logistic regression                                     Number of obs   =        294
                                                        LR chi2(3)      =       5.62
                                                        Prob > chi2     =     0.1318
Log likelihood = -145.95772                             Pseudo R2       =     0.0189

------------------------------------------------------------------------------
       drink | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .4610127   .1592889    -2.24   0.025     .2342104    .9074437
   depressed |   .6436782   .5415789    -0.52   0.601     .1237324    3.348528
    depbysex |   2.556483   2.448818     0.98   0.327     .3911017    16.71076
------------------------------------------------------------------------------
```

(d) The printout for the continuous version of the depression variable is shown below. In general an interaction means that the relationship between the outcome and one predictor depends on the value of the other predictor. Here that would mean that the difference in odds of regular drinking between men and women depends on the level of depression or equivalently that effect of the continuous depression score on likelihood of drinking is different for men versus women–i.e. the slopes are different on the log odds scale. Our log odds model is

$$ln(odds) = 2.13 + -1.12sex - .04cesd + .057sex * cesd$$

The easiest way to understand the coefficients is to look at the resulting fits for men (sex=0) and women (sex=1). For men we have

$$ln(odds) = 2.13 - 1.12(0) - .042cesd + .057(0)cesd = 2.13 - .042cesd$$

Thus we see that the log odds of being a regular drinker for a man with no depression is 2.13 (which corresponds to a probability of being a regulardrinker of $e^{2.13}/(1 + e^{2.13}) = .89$ or 89% and for every additional point more depressed on the cesd scale the log odds go down by .04 (odds ratio $e^{-.042} = .96$ or a 4% decrease in odds per point of depression. (These numbers may seem a bit odd but of course it is not clear whether this slope is actually different from 0.) For women the equation is

$$ln(odds) = 2.13 - 1.12(1) - .042cesd + .057(1)cesd = 1.01 + .015cesd$$

For a woman with no depression the log odds of being a regular drinker are 1.01 (correpsonding probability 73.3%) and the log odds go up by .015 for each additional point on the depression scale (odds ratio 1.015 or a 1.5% increase in odds per point of depression.)

Overall it seems like women have a lower baseline rate of regular drinking but their odds of being a regular drinker may increase with depression while men's odds of drinking may go down with increased depression. However we really can't read much into these interpretations because the interaction term in this model isn't significant (p-value .135). Its p-value is smaller than that of the interaction term in the model with categorical depression but not enough that we can be sure of our results. A sketch of the log likelihood equations derived above is shown in the accompanying graphics file.

```
. logit drink sex cesd cesdbysex

Logistic regression                                     Number of obs   =        294
                                                        LR chi2(3)      =       6.25
                                                        Prob > chi2     =     0.1000
Log likelihood = -145.64143                             Pseudo R2       =     0.0210

------------------------------------------------------------------------------
       drink |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -1.115571   .4703699    -2.37   0.018    -2.037479   -.1936627
        cesd |  -.0418626    .032501    -1.29   0.198    -.1055633    .0218382
   cesdbysex |    .056568   .0378018     1.50   0.135    -.0175222    .1306582
       _cons |   2.127892   .4010354     5.31   0.000     1.341877    2.913907
------------------------------------------------------------------------------
```

(5) Homework Is Depressing:

(a) The printout for the model with all four variables is shown below. All four of the predictors have significant p-values so seem to be associated with depression. The coefficient for sex is positive and its odds ratio is above 1, indicating that all else equal females are more likely to be depressed than males which is consistent with conventional wisdom. The coefficient for age is negative and the odds ratio is less than 1 meaning that, all else equal, as your age increases your likelihood of being depressed decreases. It is less clear whether this is in the expected direction. In some senses increased ses, confidence, maturity and so on come with age and might be expected to decrease depression but eventually loss of the social network and ill health can contribute to increased risk of depression. Here we are adjusting for some of those factors and the net effect seems to be that all else equal older people will be less depressed but of course there are other age-related factors we haven't adjusted for or a curvilinear relationship between age and depression might be more appropriate. Income has a negative coefficient/odds ratio below 1 so higher income is associated with lower risk of depression which is definitely consistent with what we would expect. Health has a positive coefficient and an odds ratio above 1 meaning that a higher score on the health scale is associated with a higher risk of depression. At first this may sound backwards but the health variable is coded as 0 = excellent and 3 = poor so in fact we see that poor health is associated with increased risk of depression as expected–being in poor health physically does not improve your emotional state!

```
. logit depressed sex age income health
```

```
Logistic regression                               Number of obs   =        294
                                                  LR chi2(4)      =      28.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -119.84674                       Pseudo R2       =     0.1060

------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .8937838   .3903319     2.29   0.022     .1287474     1.65882
         age |  -.0289694   .0097782    -2.96   0.003    -.0481343   -.0098045
      income |  -.0333487    .013999    -2.38   0.017    -.0607863   -.0059111
      health |    .549148   .1947193     2.82   0.005     .1675051    .9307909
       _cons |  -.8516017   .5863836    -1.45   0.146    -2.000892    .2976889
------------------------------------------------------------------------------

. logistic depressed sex age income health

Logistic regression                               Number of obs   =        294
                                                  LR chi2(4)      =      28.43
                                                  Prob > chi2     =     0.0000
Log likelihood = -119.84674                       Pseudo R2       =     0.1060

------------------------------------------------------------------------------
   depressed | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.444361    .954112     2.29   0.022     1.137403     5.25311
         age |   .9714462    .009499    -2.96   0.003     .9530058    .9902434
      income |   .9672012   .0135399    -2.38   0.017     .9410243    .9941064
      health |   1.731777   .3372104     2.82   0.005     1.182351    2.536514
------------------------------------------------------------------------------
```

(b) The model with the indicators for the health categories is shown below. I used 0 (excellent health) as the reference so I don't use it's indicator in the model–its effective coefficient in this model is 0. (The overall intercept is setting the baseline for this group but we don't have to include it in the plot because we are really plotting relative change from excellent health.) To check linearity we plot the coefficient for the indicator versus the midpoint of the bin with which it is associated. Here since health can only take on the values 0,1,2 or 3 there are no "bins" to take midpoints of–we just use the actual values. The plot is shown in the accompanying graphics file. To create it I manually entered two new columns in my data set–one consisting of the coefficients (with 0 for the reference group) and one consisting of the health levels (0-3). The points come very close to following a straight line so the assumption of a linear relationship between health and the log odds of depression seems justified. Another way of assessing this is to look at whether we did any better by allowing separate indicators for each level of health than just using the single continuous variable. We see that the log likelihood for the model with the indicators is -119.40252, nearly identical to the log likelihood of -119.84674 that we got from the model with the single continuous health variable. Allowing an arbitrarily flexible health relationship did not improve our fit. This also supports the idea that the linear relationship is adequate. (Which means we don't have to test a new model–don't say I never give you a break!!) In fact, although it may not look like it at first, the model with the continuous health variable is nested in the model with the indicators for the health levels–linearity just constrains the coefficients of the indicators to have a specific relationship to one another. Thus we could actually do a likelihood ratio chi-squared test to see if the model with the indicators was a significant improvement but here it is totally obvious that it is not.

```
. logit depressed sex age income health1 health2 health3
```

```
Logistic regression                                     Number of obs   =        294
                                                        LR chi2(6)      =      29.32
                                                        Prob > chi2     =     0.0001
Log likelihood = -119.40252                             Pseudo R2       =     0.1094

------------------------------------------------------------------------------
   depressed |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   .9162367   .3923203     2.34   0.020     .1473031     1.68517
         age |  -.0293057    .009947    -2.95   0.003    -.0488014   -.0098099
      income |  -.0341426    .014046    -2.43   0.015    -.0616723   -.0066129
     health1 |   .2461406   .3859411     0.64   0.524    -.5102901    1.002571
     health2 |   .9850562   .4980601     1.98   0.048     .0088764    1.961236
     health3 |    1.79836   .6764804     2.66   0.008     .4724828    3.124237
       _cons |  -.7145986   .6006344    -1.19   0.234     -1.89182    .4626231
------------------------------------------------------------------------------
```

(c) The Hosmer-Lemeshow test basically bins the observations (usually by probability of an event) and checks whether the observed fraction of cases in each bin matches the fraction of cases in each bin predicted by the model. If the observed proportions match the "expected" proportions from the model then the model fits well. If there are big differences between the observed and expected proportions then the model does not fit will. The summary statistic for the H-L test is a chi-squared statistic analogous to that used in a standard contingency table analysis and a large test statistic/significant p-value suggests there is a significant difference between the model and the observed data–i.e. if the p-value is small you reject the hypothesis of "no difference" between the observed and expected values and conclude the model does **not** fit well. The printout for the HL test for this model is shown below. Since some of our variables are continuous I used the version where you create bins by ordering the observations from lowest to highest predicted probabilities and dividing them into 10 equal groups or deciles. The p-value here is far from significant at .4132, suggesting that this model fits fairly well. Note however that the goodness of fit does NOT have to improve in this sense as you add more variables. If you add in a new variable but get the shape of its relationship with the outcome wrong you may actually make the calibration worse!

```
. estat gof, group(10)

Logistic model for depressed, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =       294
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =      8.21
                Prob > chi2 =      0.4132
```

(d) The pseudo-R^2 based on the log likelihood is given in the basic logistic printout. It is only .106 for this model, indicating that despite the enormous overall significance of the model and the individual predictors, and the fact that the model fits quite well (per the HL test), we are still a long way from a perfect fit or perfect predictions. If you like to think of it this way we have moved only 10.6% of the way along the path from the log-likelihood of the null model (no predictors) to the saturated model (perfect fit.) I mentioned other pseudo-R^2 measures in class. One is the correlation between the Y values and the predicted probabilities or

(if you want to be on the same scale as the other pseudo-R^2 values, the square of that correlation. Below I give the commands for obtaining the predicted probabilities and the correlation. The resulting pseudo-R^2 is $r^2 = .3033^2 = .092$ which is slightly worse than the .106 we get from the likelihood approach. Another option is to use sums of squares, specifically

$$1 - \frac{\sum(y_i - \hat{p}_i)^2}{\sum(y_i - \bar{y})^2} 2 = 1 - \frac{.128}{.141} = .092$$

I got the values for the sums of squares by creating two new variables, the squared differences between the Y's and the predicted probabilities based on the fitted model and the squared differences based on the null model (which just predicts the overall fraction of the cases in the sample as the probability for everyone). The ratio of the means for these variables is the same as the ratio of the sums of squares since the sample size is the same in each case. The corresponding output is shown below. This metric produces about the same pseudo-R^2 as the correlation approach. The main point here is to recognize that although the model fits well and is significant there is still a lot of unexplained variability–this is very typical of logistic regression. Moreover there are many different measures of R^2 and they are not completely consistent in their numerical values. These measures are most useful for comparing different models with the same outcome variable to see how much better one performs than another.

```
Note that the following commands must be used right after fitting the
relevant logistic regression model!

. predict depprobs
(option p assumed; Pr(depressed))
. gen rawresids = depressed - depprobs
*********
Correlation pseudo R-squared


. cor depressed depprobs
(obs=294)

             | depres~d depprobs
-------------+------------------
   depressed |   1.0000
    depprobs |   0.3033   1.0000

************************************
Sum of squares pseudo R-squared
. sum depressed

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   depressed |       294     .170068    .3763331          0          1

. gen depresidsnull2 = (depressed - .17)^2

. gen rawresids2 = rawresids^2

. sum depresidsnull2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
depresidsn~2 |       294    .1411449    .2483799      .0289      .6889

. sum rawresids2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  rawresids2 |       294     .128179    .2223346   .0001713   .9076856
```

(e) and (f) The desired sensitivity, specificity and total fraction correct plots are shown in the accompanying graphics file. (I did not require you to obtain the last of these but I include it for reference as it is often useful.) The sensitivity and specificity are calculated easily by our computer packages. The total fraction correct can be calculated from the sensitivity and specificity as follows. Let n_P be the number of positive cases in the sample and n_N is the number of negative cases with $n = n_P + n_N$. Let n_{TP} be the number of positive cases that are predicted correctly (the true positives) and let n_{TN} be the number of negative cases predicted correctly (the true negatives). The sensitivity is n_{TP}/n_P and the specificity is n_{TN}/n_N. The fraction correct is

$$\frac{n_{TP} + n_{TN}}{n} = \frac{n_P(n_{TP}/n_P) + n_N(n_{TN}/n_N)}{n} = \frac{n_P}{n} Sensitivity + \frac{n_N}{n} Specificity$$

The commands for generating this variable are also shown below. The sensitivity/specificty plot shows the two curves crossing at a threshold of about .17 and the corresponding sensitivity and specificity (and hence total fraction correct) are both around 66% which isn't bad.

The corresponding area under the ROC curve is .73 which is also OK though not spectacular. Overall this model performs well in terms of predictive accuracy, despite the rather low seeming pseudo-R^2 values. Note that STATA has a cool command called **estat class** with an option **cutoff** which will give you the cross-classification table, sensitivity, specificity and pretty much every other summary you can think of for a particular probability threshold. I experimented a little to get the point where sensitivity and specificity were about equal for these data but eyeballing the plot was fine. The printout is shown below.

```
. lsens, gensens(depsens) genspec(depspec) genprob(depthreshold)

. tab depressed

  depressed |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        244       82.99       82.99
          1 |         50       17.01      100.00
------------+-----------------------------------
      Total |        294      100.00

 gen totcorrect = .1701*depsens + .8299*depspec
(718 missing values generated)

. scatter totcorrect depthreshold

. lroc
```

```
Logistic model for depressed

number of observations =      294
area under ROC curve   =   0.7314
************************************************************
. estat class, cutoff(.17)

Logistic model for depressed

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        33            82  |        115
     -     |        17           162  |        179
-----------+--------------------------+-----------
   Total   |        50           244  |        294

Classified + if predicted Pr(D) >= .17
True D defined as depressed != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   66.00%
Specificity                     Pr( -|~D)   66.39%
Positive predictive value       Pr( D| +)   28.70%
Negative predictive value       Pr(~D| -)   90.50%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   33.61%
False - rate for true D         Pr( -| D)   34.00%
False + rate for classified +   Pr(~D| +)   71.30%
False - rate for classified -   Pr( D| -)    9.50%
--------------------------------------------------
Correctly classified                        66.33%
--------------------------------------------------
```