Solutions To Homework Assignment 1 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) Maximum Likelihood Basics:

(a) Likelihood Function: The likelihood function represents how likely your data are to have occurred given a particular generating distribution and set of parameters. The function is obtained by multipling the probability mass function (for a discrete Y variable) or probability density function (for a continuous Y variable) values for each of the points in your data set. For example, for the normal distribution, the probability density function is

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(Y-\mu)^2}$$

This function gives the relative "likeliness" of different values of Y assuming Y is normally distributed with mean μ and variance σ^2. The function has a bell-shape with it's peak at μ, tailing off symmetrically to either side. In a standard regression model the mean, μ, depends on the values of the predictor variables, X, with $\mu = \beta_0 + \beta_1 X_1 + \cdots \beta_m X_m$. If we have sample of size n for which we have measured the outcome, Y, and the predictors, $X_1, \ldots, X_m$ then our likelihood for the whole data set becomes (with some abuse of notation)

$$\Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(Y_i - X_i\beta)^2}$$

(b) Maximum Likelihood Estimation: The intuitive idea behind maximum likelihood estimation is that you want to pick as your parameter estimates those values that are "most likely" to have generated your observed data. These in some natural sense represent the best fit of your model to your data. To do this mathematically one writes down an expression for the likelihood of one's data under the proposed model (e.g. the likelihood equation above represents a standard linear regression model) as a function of the model parameters (in this case the regression coefficients, $\beta_0, \beta_1, \ldots, \beta_m$ and σ^2). One then differentiates the likelihood with respect to each of the parameters and sets the results equal to 0 to obtain a system of equations. (The maximum of any continuous differentiable function occurs at a place where the derivative is 0 since the slope is changing from positive–i.e. going up the hill–to negative–i.e. going down the hill.) The solution to this system of equations gives the maximum likelihood estimates for the parameters. In linear regression this is the same as the least squares solution since the equation above is a function of the squared difference between the Y values and the values predicted by the model ($X\beta$) and one can write down an explicit formula for the MLE for each β. However in a logistic regression the likelihood is messier and there is no closed form formula for the coefficient estimates.

(c) MLE for a Normal, Part I: With a sample of size n=1 and no predictor variables our expression for the normal likelihood is the first formula given above. This function is simply the bell-curve which has its peak at μ. Since our observed value is Y=10 we will get the maximum value of the likelihood by choosing $\mu = 10$. This should intuitively make sense. To see it mathematically we note that the exponential term includes a negative sign so the bigger the piece $(Y - \mu)^2$ is, the smaller the likelihood will get. The best we can do is to make this 0 by choosing $Y = \mu$. We can not however get a reasonable value for the MLE of σ^2 from this data set. Intuitively this is because we only have one data point so there is no variability and σ^2 is the variance of the distribution. One could argue that the variance in our data set is 0 but mathematically this makes the likelihood function undefined.

(d) MLE for a Normal, Part II: Of course you're never going to have a sample of size n=1 (at least I hope not!) If you have a bigger sample, not all the values will be the same and so the choice of μ will have to represent a compromize among the observed values. Not surprisingly the MLE turns out to be the sample mean, $\bar{Y}$. To see this we note that the likelihood, assuming all n observations are drawn from a population with the same mean and variance, is

$$\Pi_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(Y_i-\mu)^2}$$

To maximize this it is easiest to first take the natural log which turns the product into a sum and gets rid of the exponentials. Maximizing the log likelihood is equivalent to maximizing the likelihood since the log is a monotone increasing function. The constant term before the exponential doesn't depend on μ so we can ignore it for purposes of getting the MLE for the mean. We just need to maximize

$$\sum_{i=1}^{n} -\frac{1}{2\sigma^2}(Y_i - \mu)^2$$

Differentiating this with respect to μ and setting to 0 gives us

$$\frac{-1}{2\sigma^2} \sum (Y_i - \mu) * (-1) = 0$$

which after a little algebra yields $\sum Y_i = n\mu$ or $\mu = \bar{Y}$. Note that this is a bit of an abuse of notation since the true mean, μ isn't really the sample mean. People often write the likelihood function using a Roman rather than a Greek letter for the parameters or else put a "hat" on the parameter symbols to reflect that you are finding the best estimate.

(2) Generalized Linear Model Basics: (a) The three basic components of a generalized linear model are (i) the distribution of Y, (ii) a systematic component represented by a linear combination of the predictor variables, $X\beta = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$ and (iii) a link function, $g(\mu)$ which relates the average value of Y to the systematic component: $g(E(Y|X's)) = g(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m$. The regression coefficients, the β's are estimated using maximum likelihood as described in Problem 1. In some special cases there is a closed form solution for the MLEs but in general there is not and the actual fitting procedure involves an iterative optimization algorithm such as iteratviely reweighted least squares.

(b) The formula for the **deviance** is

$$D = -2ln(\frac{L_{full}}{L_{saturated}}) = -2(ln(L_{full}) - ln(L_{saturated}))$$

where L represents the likelihood of the specified model. Essentially the deviance compares (in ratio form) the likelihood for the specified model and a model which allows a separate parameter for each data point, leading to (as close as one can get to) a perfect fit. Taking the logarithm and multiplying by -2 standardizes this ratio so that it has a chi-squared distribution. Intuitively the deviance is like SSE in a standard linear regression. It represents how much worse your model is than one that provides a perfect fit. That model with

a "perfect fit" is called the **saturated model** because it is "completely stuffed"–there is one parameter for each data point and you couldn't possibly estimate any more parameters. The **null model** in constrast is the one that has no predictor variables. It represents the worst cases scenario where you have no information except the observed outcome values for a sample of size n. Together these two models provide the anchors for measuring how well your actual or **full** model (the one including your X variables) is fitting. It can't do any worse then the null model or any better than the saturated model. In the standard regression context, the null model corresponds to guessing $\bar{Y}$ as the predicted value no matter what your X values are and the corresponding error is SST, the sum of squares total. The error for the full model, taking advantage of the X variables, is SSE. The error for the saturated model is 0–you have fit each data point exactly. The deviance is a generalization of SSE–it tells you how far you are from the best possible fit.

(3) Logistic Regression Basics:

(a) For a logistic regression, the distribution of Y is assumed to be Binomial with n=1 (or if you prefer, Bernoulli) and a success probability, p, which depends on a set of predictors or X's. The systematic part is, as always, just the linear combination of those predictors, $X\beta$. The expected value of a Binomial distribution is just the probability, p, since a Binomial distribution with n=1 takes on only the values 1 and 0 with probabilities p and 1-p respectively. (In general $\mu_Y = E(Y) = \sum_y y * P(Y = y)$ which for a Binomial with n = 1 is just $1 * p + 0 * (1 - p) = p$.) Finally, the link function, relating the probability of of the event of interest to the systematic component is $g(p) = ln(p/(1-p))$, known as the **logit function**. The logit function is chosen to transform the probability, which must lie between 0 and 1, to a number that can take any real value. This is necessary since the systematic component, $X\beta$ can take on any value between negative and positive infinity. The logit function is actually what is called the **canonical link** for the binomial distribution which means that it mathmatically arises naturally out of the likelihood function for the binomial distribution.

(b) The coefficients in a logistic regression provide the same information as they do in a standard linear regression except that they must be interpreted on the log odds scale. Specifically, for a continuous variable, X, the coefficient β gives the change in **log odds** associated with a one unit change in X, all other variables held fixed. For an indicator variable, β gives the difference in log odds between subjects who do and do not have the characteristic coded for by the indicator.

(c) The log odds scale is not very intuitive so it is natural to start transforming the regression coefficients back towards the probability scale. Unfortunately because of the non-linearity of the logit transformation it is hard to give a nice interpretation on the probability scale. The change in probability associated with a particular change in X depends on what the value of X was initially. However we can give a reasonable interpretation on the odds or odds ratio scale.

What is an odds ratio? Specifically, suppose you have two subjects, A and B, and you are interested in comparing how relatively likely it is that each will experience a particular event. Let p_A be the probability of the event for the first subject and p_B be the probability of the event for the second person. Then the odds ratio is defined as

$$OR_{AvsB} = (\frac{p_A}{1 - p_A})/(\frac{p_B}{1 - p_B})$$

This is, as its name suggests, the ratio of the odds that A experiences the event to the odds that B experiences the event. An odds ratio of 2 means the odds are twice as high that A will experience the event as that B will experience the event. It is important that you be careful with your wording. The odds ratio does not give the ratio of the probabilities–make sure you word your description in terms of odds rather than "probability" or "chance" or "risk."

In a logistic regression setting, if you exponentiate the regression coefficient or β's you get an odds ratio telling you how a change the particular X variable is related to (but don't think causal effect here!) the odds

of an event. Specifically, for a continuous variable, the odds ratio of the coefficient gives you the (multiplicative) change in odds associated with a one unit change in X. To make this concrete, consider our CHD and age example from class. The odds ratio of 1.12 obtained by exponentiating the coefficient of the age variable tells us that each extra year of age is associated with 1.12 times higher odds or an increase of 12% in the odds of a person having coronary heart disease. Another way to say this is that if I have two people who are one year apart in age, the older one will have 12% higher odds of having CHD than the younger one, all else equal. The odds ratio for a categorical variable is actually a little simpler since in this case you are just comparing the odds for people who have a characterstic to those who don't. For example, I did a version of the CHD data where I dichotomized the age variable into "below 50" or above 50". The odds ratio corresponding to this age group variable would simply tell us the relative odds of CHD in people over and under 50. From this model the odds ratio was over 8, meaning that people over 50 years old had odds more than 8 times as high of having coronary heart disease as people under 50.

(4) Sports Fantatics:
(a) There is evidence that at least one of the variables is a significant predictor of whether or not the All Blacks win. The p-value for the overall likelihood ratio chi-squared test is very small (< 0.0001). This indicates that we can reject the null hypothesis that none of the variables are helping to predict wins. At least one variable is a significant predictor. Mathematically our hypotheses would have been $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$ versus H_A : At least one $\beta \neq 0$.

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

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

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

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

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

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

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

Here we have b = .115, Z = 1.96 for a 95% confidence interval, $s_b = .045$ and $\Delta = 10$. Plugging these numbers in gives a CI of [1.307, 7.629]. Thus the odds of winning are somewhere between 1.307 and 7.629 times as high if the temperature goes up 10 degrees, all else being equal. This is a very wide range so we don't have much precision!

(c) AB Win% (p-value 0.0082), Opp Win% (p-value 0.0081), Home? (p-value 0.0278) and Temperature (p-value 0.0108) are all statistically significant variables because they have low p-values. The Australia indicator is not significant because its p-value is above $\alpha = .05$. As far as the signs, the better the All Blacks have been

playing the more likely they are to keep winning so the positive sign on AB Win% makes sense. However if the All Black's opponent has been playing well it will be a harder game so the chances of winning will go down. Thus the negative sign on Opp Win% makes sense. Similarly home field is an advantage so we would expect the Home? coefficient to be positive as it is. New Zealand is a warm country so it is not surprising the All Blacks play better in warmer weather as suggested by the positive sign on the temperature variable. The All Black's archrival Australia is the 2nd best team in the world (compared to the All Black's of course!) so games against them are harder and we would expect a negative coefficient. The fact that this isn't significant indicates just how good the All Blacks are! Of course I wouldn't expect you to know these extra rugby facts....

(d) The formula for the predicted probability is

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

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

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

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

(5) Asthma As Math: This problem parallels the turn-in problems very strongly so it is a useful reference!

(a) We are asked to perform a two-sample test of proportions to see whether there is a higher rate of asthma in children who lived in an urban setting as infants than in those who did not. The test is one-sided because Dr. Green wants to show that an urban environment is associated with increased risk. If we let p_U be the probability that a child born in an urban environment gets asthma and p_R be the probability that a child born a rural environment develops asthma then our hypotheses are

$H_0 : p_U \leq p_R$–the asthma rate in an urban environment is the same or even lower than that in a non-urban area.
$H_A : p_U > p_R$–an urban environment is associated with an increased risk of a child developing asthma.

The test statistic for evaluating these hypotheses is based on the difference in sample proportions between the two groups. Specifically our test statistic is

$$Z = \frac{(\hat{p}_U - \hat{p}_R) - 0}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_U}+\frac{1}{n_R})}}$$

Where $\hat{p}$ is the proportion of children in the combined sample who got asthma. Note that the boundary value under the null hypothesis is that the proportions are the same in the two groups so it makes sense to

use a pooled estimate for our standard error. Under the null hypothesis this test statistic has a standard normal distribution. The printout below shows the two sample proportions, the value of the Z statistic, and the corresponding p-values for both the one-sided and two-sided tests. I have set this up as a one-sided test by asking whether there was in increased risk in the urban group. If you just wanted to know whether the rates were different you would do a two-sided test. We see that 8.25% of children in the rural group developed asthma while 15% of children in the urban group developed asthma. The corresponding p-value for the one-sided test is .0007. This is very small so we reject the null hypothesis and conclude there is a higher risk in the urban group.

We are also asked to calculate the odds ratio comparing the relative likelihood of developing asthma in urban-born vs rural-born children. In terms of the values p_U and p_R the required odds ratio is

$$OR = (\frac{p_U}{1-p_U})/(\frac{p_R}{1-p_R}) = \frac{.15/.85}{.0825/.9175} = 1.96$$

The estimated value was obtained by plugging in the sample proportions for the two groups. It looks as if the odds of developing asthma are nearly twice as high for city children as for rural children.

```
. prtest asthma, by(urban)

Two-sample test of proportion                      0: Number of obs =      400
                                                   1: Number of obs =      600
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |      .0825   .0137562                      .0555382    .1094618
           1 |        .15   .0145774                      .1214289    .1785711
-------------+----------------------------------------------------------------
        diff |     -.0675   .0200433                     -.1067842   -.0282158
             |  under Ho:   .0212005    -3.18   0.001
------------------------------------------------------------------------------
        diff = prop(0) - prop(1)                                      z =  -3.1839
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.0007         Pr(|Z| < |z|) = 0.0015          Pr(Z > z) = 0.9993

IN SAS ANALYST:
Two Sample Test of Equality of Proportions

Sample Statistics
- Frequencies of asthma for urban -

Value         0             1
 ----------------------------------------
0           367           510
1            33            90

Hypothesis Test
Null hypothesis:
Proportion of asthma(urban=0) - Proportion of asthma(urban=1) =  0
```

```
Alternative:
Proportion of asthma(urban=0) - Proportion of asthma(urban=1) ^= 0

                - Proportions of asthma for urban -
 Value          0             1                  Z     Prob > Z
-------------------------------------------------------------------
1                    0.0825        0.1500    -3.18    0.0015
```

(b) We can also view these data as coming from a 2x2 contingency table. A contingency table records how many people there are with each of the possible combinations of values of two categorical variables. Here we have four possibilities–the child was born in a city and got asthma, was born in a city and didn't get asthma, was born outside a city and got asthma or was born outside a city and didn't get asthma. Under the null hypothesis assumption that there is no association between whether or not you were born in a city and whether or not you develop asthma we can calculate the "expected" number of people in each of these categories. For example if 60% of people overall are born in a city and 25% of children overall develop asthma (totally made up numbers!) then we would expect 15% of kids to be born in cities and develop asthma, 45% to be born in cities and not develop asthma, 10% of kids to be born outside of cities and develop asthma and 30% to be born outside of cities and to not develop asthma. Of course if children born in cities are more likely to develop asthma then we will have more city-asthma cases than expected and fewer rural-asthma cases than expected. This can be formalized by performing Pearson's chi-squared test of independence. The test statistic quantifies the difference between the observed and expected number of people in each of the categories. The contingency table and corresponding chi-squared statistic and p-value for this data set are shown below. Note that I got STATA and SAS to show me the expected number of people in each cell as well as the actual numbers. We see that as we suspected we got more urban asthma cases (90) than expected (73.8) based on the overall numbers of asthma cases and urban dwellers and correpsondingly fewer rural asthma cases than expected. The chi-squared statistic was highly significant at .001 so these data suggest a relationship between asthma development and living environment. Note that the the chi-squared test simply evaluates whether or not there is a relationship which is a two-sided test. To convert this to a 1-sided p-value we would need to divide the p-value from the printout in half. However even the two-sided test is significant!

In the case of two groups with two possible outcomes, the chi-squared test is completely mathematically equivalent to the 2-sample test of proportions and in fact the chi-squared statistic is just the square of the Z statistic. We can verify this for our data. From (a) we got Z = -3.18 and from (b) we got $\chi^2 = 10.14 = (3.18)^2$ up to rounding error.

IN STATA:

```
. tab asthma urban, chi2 exp

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
+--------------------+

           |         urban
    asthma |         0          1 |     Total
-----------+----------------------+----------
         0 |       367        510 |       877
```

```
           |     350.8      526.2 |     877.0
-----------+----------------------+----------
         1 |        33         90 |       123
           |      49.2       73.8 |     123.0
-----------+----------------------+----------
     Total |       400        600 |     1,000
           |     400.0      600.0 |   1,000.0

          Pearson chi2(1) =  10.1371   Pr = 0.001
************************************************
IN SAS:

proc freq data = tmp1.hw1;
table asthma*urban/chisq;
run;

Note: I cut the printed table because it didn't cut and paste nicely into
my mathematical word processor. It is the same as STATA's. The various
tests however are given below.

                                The FREQ Procedure

                        Statistics for Table of asthma by urban

                    Statistic                     DF       Value      Prob

                    Chi-Square                     1     10.1371    0.0015
                    Likelihood Ratio Chi-Square    1     10.6025    0.0011
                    Continuity Adj. Chi-Square     1      9.5210    0.0020
                    Mantel-Haenszel Chi-Square     1     10.1270    0.0015
                    Phi Coefficient                       0.1007
                    Contingency Coefficient               0.1002
                    Cramer's V                            0.1007


                           Fisher's Exact Test

                      Cell (1,1) Frequency (F)       367
                      Left-sided Pr <= F          0.9996
                      Right-sided Pr >= F      8.280E-04

                      Table Probability (P)    4.280E-04
                      Two-sided Pr <= P           0.0016

                             Sample Size = 1000
```

(c) Viewing this as a logistic regression we are interested in whether the regression coefficient for the urban indicator is significant and postive since that would correspond to an increased risk of asthma for that group. Our hypotheses are

$\beta_1 \leq 0$–children born in an urban environment have the same or lower risk of developing asthma than children born in a rural environment.
$\beta_1 > 0$–the probability of developing asthma is higher for children born in an urban setting than for those born in a rural setting.

The printout is shown below and once again we see a significant relationship. The coefficient of the urban indicator is .67 which is positive, indicating that the urban environment is associated with higher risk. The p-value for a two-sided test is either .0011 (using the likelihood ratio chi-squared test) or .002 using the Wald test. To get the one-sided p-value we divide the two-sided p-value in half, yielding .0055 or .001. Either way we conclude that the risk is significantly higher in the urban group.

Finally we are asked about the relationship between the sample proportions and the regression coefficients. The intercept in a logistic regression is the log odds of the event for subjects whose predictor variables are all 0. Here $X = 0$ corresponds to a child born in a rural environment. Thus the intercept must be estimated by

$$b_0 = ln(\frac{\hat{p}_R}{1-\hat{p}_R}) = ln(.0825/.9175) = -2.41$$

which is what we got from the logistic model. Similarly, the coefficient of the predictor variable in a logistic regression is the log of the odds ratio for that variable. In part (a) we found the odds ratio estimate was 1.96 so the estimated regression coefficient must be

$$b_1 = ln(OR) = ln(1.96) = .67$$

again consistent with the printout.

```
IN STATA:

 logit asthma urban

Iteration 0:   log likelihood = -372.85997
Iteration 1:   log likelihood = -367.64865
Iteration 2:   log likelihood = -367.55878
Iteration 3:   log likelihood = -367.55872

Logistic regression                                     Number of obs   =       1000
                                                        LR chi2(1)      =      10.60
                                                        Prob > chi2     =     0.0011
Log likelihood = -367.55872                             Pseudo R2       =     0.0142

------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   .6742532   .2147084     3.14   0.002     .2534326    1.095074
       _cons |  -2.408854   .1817355   -13.25   0.000    -2.765049   -2.052659
------------------------------------------------------------------------------

. logistic asthma urban

Logistic regression                                     Number of obs   =       1000
                                                        LR chi2(1)      =      10.60
```

```
                                                Prob > chi2     =     0.0011
Log likelihood = -367.55872                     Pseudo R2       =     0.0142

------------------------------------------------------------------------------
      asthma | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   1.962567   .4213795     3.14   0.002     1.288441    2.989404
------------------------------------------------------------------------------
*******************************************************************************

IN SAS:
proc logistic data = tmp1.hw1 desc;
model asthma = urban;
run;
```

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4089	0.1817	175.6876	<.0001
urban	1	0.6743	0.2147	9.8616	0.0017

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
urban	1.963	1.288	2.989

(d) The proportions of subjects with asthma in each pollution bin are shown below. We see that the fraction of asthma cases goes steadily up as the pollution rises, consistent with what we would expect. The figure is shown in the accompanying graphics file. We see that for low levels of pollution (e.g. bin 1) the asthma rate is almost non-existent; by the time the average pollution is in the 15-20 range (bin 4) the asthma rate is around 20%. For most of this range it seems the rate is going fairly steadily up but then there seems to be a big jump at the high end (pollution levels 35-40). This will probably be smoothed out a bit by the logistic model. Since the proportion of asthma cases is steadily increasing we expect that there will be a significant (positive) relationship between pollution level and probability of developing asthma. This is confirmed by the highly significant p-values for the chi-squared test and Wald test in the logistic regression printout below.

```
Bin   Range   Asthma%
1      0-5         0
2      5-10     .077
3     10-15     .063
4     15-20     .100
5     20-25     .124
6     25-30     .170
7     30-35     .168
8     35-40     .389
**************************************
IN STATA:

. logit asthma pollution
Iteration 0:   log likelihood = -372.85997
Iteration 1:   log likelihood = -360.97073
Iteration 2:   log likelihood = -360.59228
Iteration 3:   log likelihood =  -360.5918

Logistic regression                               Number of obs   =       1000
                                                  LR chi2(1)      =      24.54
                                                  Prob > chi2     =     0.0000
Log likelihood =  -360.5918                       Pseudo R2       =     0.0329
------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   pollution |   .0682789   .0141742     4.82   0.000     .0404979    .0960598
       _cons |  -3.487279   .3476386   -10.03   0.000    -4.168638   -2.805919
------------------------------------------------------------------------------

. logistic asthma pollution
Logistic regression                               Number of obs   =       1000
                                                  LR chi2(1)      =      24.54
                                                  Prob > chi2     =     0.0000
Log likelihood =  -360.5918                       Pseudo R2       =     0.0329
------------------------------------------------------------------------------
      asthma | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   pollution |   1.070664   .0151758     4.82   0.000     1.041329    1.100825
------------------------------------------------------------------------------
*******************************************************************************
IN SAS:
proc logistic data = tmp1.hw1 desc;
model asthma = pollution;
run;

                                The LOGISTIC Procedure

                        Analysis of Maximum Likelihood Estimates

                                             Standard          Wald
             Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq
```

Intercept	1	-3.4873	0.3476	100.6266	<.0001
pollution	1	0.0683	0.0142	23.2042	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
pollution	1.071	1.041	1.101

(e) The printout for the model with urban and pollution as predictors is shown below. We see that the pollution variable remains significant (Wald test Z = 3.7 with a corresponding p-value of 0) but that the indicator for urban has become extremely non-significant and in fact the sign on the coefficient has become negative, implying that if anything after adjusting for pollution level an urban environment is actually lower risk. Of course we can't take that too seriously since it isn't significant. I suspect that there is some sort of mediation going on here. If the main reason that cities have high asthma rates is that they have more pollution then once we account for the pollution levels the city effect would go away. We could formally check whether the data are consistent with a mediation hypothesis by also testing whether there is a difference between urban and rural pollution levels in this data–and in fact there is. However do keep in mind that just because the data are consistent with a mediation hypothesis doesn't mean that have established a causal path....

IN STATA:

```
logit asthma urban pollution                      Number of obs   =       1000
                                                  LR chi2(2)      =      24.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -360.57094                       Pseudo R2       =     0.0330
------------------------------------------------------------------------------
      asthma |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |  -.0597649   .2923861    -0.20   0.838    -.6328312    .5133013
   pollution |   .0709072   .0191487     3.70   0.000     .0333765    .1084379
       _cons |  -3.506273   .3596161    -9.75   0.000    -4.211108   -2.801439
------------------------------------------------------------------------------
. logistic asthma urban pollution

Logistic regression                               Number of obs   =       1000
                                                  LR chi2(2)      =      24.58
                                                  Prob > chi2     =     0.0000
Log likelihood = -360.57094                       Pseudo R2       =     0.0330

------------------------------------------------------------------------------
      asthma | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       urban |   .9419859   .2754236    -0.20   0.838      .531086    1.670798
   pollution |   1.073482   .0205557     3.70   0.000      1.03394    1.114536
------------------------------------------------------------------------------
*******************************************************************************
```

```
IN SAS:
proc logistic data = tmp1.hw1 desc;
model asthma = urban pollution;
run;
```

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.5063	0.3596	95.0631	<.0001
urban	1	-0.0598	0.2924	0.0418	0.8380
pollution	1	0.0709	0.0191	13.7121	0.0002

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
urban	0.942	0.531	1.671
pollution	1.073	1.034	1.115

(f) The printouts for the two models are shown below. The likelihood ratio chi-squared test allows us to determine whether one of these models is significantly better than the other by determining whether it has a much higher likelihood of having generated our observed data. The hypotheses are

$H_0 : \beta_1 = \beta_2 = \cdots \beta_6 = 0$–none of these six predictors (urban environment, pollution, ses, breastfeeding duration, family history or sex assigned at borth) is useful for predicting the probability of developing asthma
H_A : At least one of the $\beta_j \neq 0$–at least one of these factors is related to risk of developing asthma.

The test statistic for the likelihood ratio chi-squared test can be thought of the difference between the deviances of the two models or else is -2 times the difference in log likelihood between the smaller model and the larger model. Here, using the log likelihood from the printouts we have

$$\chi^2_{obs} = -2(-372.86 - (-331.84)) = 82.04$$

Under the null hypothesis this statistic has a chi-squared distribution with 6 degrees of freedom (correpsonding to the 6 predictors in the model.) We can get the tail probability using a distributional calculator (see below) or we can simply look at the test statistic and p-value for the overall test in the second model. The chi-squared statistic matches what we got from the likelihoods and the corresponding p-value is essentially 0. Thus we reject the null hypothesis and conclude that (not surprisingly!) at least one of these predictors is associated with asthma risk.

```
. logistic asthma

Logistic regression                               Number of obs   =       1000
                                                  LR chi2(0)      =       0.00
```

```
                                                Prob > chi2     =          .
Log likelihood = -372.85997                     Pseudo R2       =     0.0000

------------------------------------------------------------------------------
      asthma | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
------------------------------------------------------------------------------
*******************************************************************************

. logit asthma urban pollution ses breastfed famhist sex

Iteration 0:   log likelihood = -372.85997
Iteration 1:   log likelihood = -336.14862
Iteration 2:   log likelihood = -331.89197
Iteration 3:   log likelihood = -331.83842
Iteration 4:   log likelihood = -331.83839

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
         sex |  -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 sex

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
         sex |   .3346326     .07291    -5.02   0.000     .2183271    .5128953
------------------------------------------------------------------------------
```

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

. display chi2tail(6, 82.04)
1.353e-15
*******************************************************************************
IN SAS:
proc logistic data = tmp1.hw1 desc;
model asthma = ;
run;
```

The LOGISTIC Procedure

Model Information

Data Set	TMP1.HW2	
Response Variable	asthma	asthma
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	1000
Number of Observations Used	1000

Response Profile

Ordered Value	asthma	Total Frequency
1	1	123
2	0	877

Probability modeled is asthma=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

-2 Log L = 745.720

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9643	0.0963	416.2271	<.0001

```
************
proc logistic data = tmp1.hw1 desc;
model asthma = urban ses pollution breastfed famhist sex;
famchartest: test ses=breastfed=famhist=sex=0;
contrast "fam characteristics" ses 1,
breastfed 1,
famhist 1,
gender 1;
run;
```

The LOGISTIC Procedure

Model Information

Data Set	TMP1.HW2	
Response Variable	asthma	asthma
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	1000
Number of Observations Used	1000

Response Profile

Ordered Value	asthma	Total Frequency
1	1	123
2	0	877

Probability modeled is asthma=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	747.720	677.677
SC	752.628	712.031
-2 Log L	745.720	663.677

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	82.0432	6	<.0001
Score	80.2418	6	<.0001
Wald	70.6021	6	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0470	1.6558	0.0008	0.9773
urban	1	-0.0772	0.3021	0.0653	0.7983
ses	1	-0.0818	0.0346	5.5965	0.0180
pollution	1	0.0766	0.0200	14.6638	0.0001
breastfed	1	0.00573	0.0103	0.3079	0.5789
famhist	1	1.1801	0.2230	27.9956	<.0001
sex	1	-1.0947	0.2179	25.2447	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
urban	0.926	0.512	1.673
ses	0.921	0.861	0.986
pollution	1.080	1.038	1.123
breastfed	1.006	0.986	1.026
famhist	3.255	2.102	5.039
sex	0.335	0.218	0.513

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
fam characteristics	4	53.0978	<.0001

Linear Hypotheses Testing Results

Label	Wald Chi-Square	DF	Pr > ChiSq
famchartest	53.0978	4	<.0001

(g) To compare the two models we can again use a likelihood ratio chi-squared test using the likelihoods from the two printouts. This time our hypotheses are

H_0 : $\beta_3 = \cdots \beta_6 = 0$–none of the family/child characteristics (ses, breastfeeding duration, family history or sex assigned at birth) explains anything about asthma risk beyond what is explained by environmental factors (urban, pollution).
H_A : At least one of the $\beta_j \neq 0$–at least one of the family/child factors contributes significant predictove power to the model.

Our test statistic, using the likelihoods from the models in parts (e) and (f) is

$$\chi^2_{obs} = -2(-360.57 - (-331.84)) = 57.47$$

and the corresponding p-value is basically 0 using the STATA calculator, noting that our test statistic should have 4 degrees of freedom since the environment plus family model has four more predictors than the environment only model. Alternatively we could have used the follow-up test command after fitting the model from part (f) or we could have saved the two model outputs and used the lrtest command. All these results are shown below. Note that the "test" command produces a Wald-like test rather than the likelihood ratio test although the results are very similar. Regardless, the model adding the family characteristics is a significant improvement over the environmental factors only model. To get the same result in SAS I could use either the test command or the contrast command. Both are given above as part of the printout for part (f).

```
. display chi2tail(4,57.47)
9.859e-12
*******************************************************************************
. test ses=breastfed=famhist=sex=0   Note: This command assumes I have
                                        just fit the 6 variable model!
 ( 1)  ses - breastfed = 0
 ( 2)  ses - famhist = 0
 ( 3)  ses - sex = 0
 ( 4)  ses = 0

           chi2(  4) =   53.10
         Prob > chi2 =    0.0000
*******************************************************************************
. lrtest ENVIRON FAMILY                  Note: This command assumes I
                                         saved the results of the two
                                         models as ENVIRON and FAMILY
                                         when I originally fit them in
                                         (e) and (f).

Likelihood-ratio test                                 LR chi2(4)  =     57.47
(Assumption: ENVIRON nested in FAMILY)                Prob > chi2 =    0.0000
```

(h) Our outcome of interest is having asthma. Risk factors are those that are associated with **higher** likelihood of getting asthma. Variables with significant positive coefficients in the log odds (logit) model are risk factors. Equivalently we could look at the odds level (logistic) printout for significant variables with odds ratios greater than 1. It appears that higher pollution levels and having a family history of asthma are

associated with higher likelihood of getting asthma. Although breastfeeding has a positive coefficient/odds ratio above 1, it is not significant so we can not say it is a risk factor (and indeed it is generally considered to be protective!) In contrast, higher socio-economic status and being female (sex = 1) appear to be protective. Note that after adjusting for the other factors the coefficient of the urban indicator is negative suggesting a protective effect but this is not even close to being significant.

(i) For a categorical variable the odds ratio compares the odds of the event of interest for people who do and do not have the particular characteristic. Here the odds ratio of 3.25 for the family history variable means that **all else equal** a child who has a family member with a history of asthma has odds of getting asthma that are 3.25 times as high as a child who does not have relatives who have had asthma. Note that "all else equal" can be thought of here as comparing two children of the same sex with the same breastfeeding history who have the same SES and live in comparable settings (urban/rural; same pollution level.) The corresponding confidence interval for the odds ratio is [2.10, 5.04] meaning that we are 95% sure that the odds of developing asthma are somewhere between 2.1 and 5.04 times as high in children with a family member with asthma as for children without such a relative. To get the regression coefficient and its confidence interval from the estimate and interval for the odds ratio we simply take natural logs, yielding an estimate of $b_{FH} = \ln(3.25) = 1.18$ and for the confidence interval [ln(2.1), ln(5.04)] = [.74, 1.62]. To reverse the process we exponentiate.

(j) For a continuous variable the odds ratio tells you about the change in odds of the event associated with a particular degree of change in the predictor. Here the odds ratio of .92 for SES means that **all else equal** for every additional point of SES the odds of a child getting asthma go down by 8%. The confidence interal is [.86, .98] meaning that we are 95% sure that the odds of developing asthma go down somewhere between 2-14% for each additional point of SES. To get the confidence interval for the odds ratio associated with a 10 point change in SES we note that the general formula for a CI for an odds ratio corresponding to a change of Δ in a continuous predictor is

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

That is we can either find the confidence intveral for a 1 unit change in X, multiply it by Δ to get the confidence interval for a Δ unit change and then exponentiate it to convert to the odds ratio scale, or equivalently we can take the interval for the odds ratio associated with a 1-unit change in X_j and then raise it to the power Δ. Since we already have the CI for the odds ratio here it's easuer just to raise it to the power 10. The resulting interval is $[.86^{10}, .98^{10}] = [.22, .82]$. This suggests that a 10 point increase in SES could be associated with anywhere between a 18-78% rediction in the odds of getting asthma.

(k) The predicted probability is given by

$$\frac{e^{b_0 + b_1 X_1 + \cdots b_6 X_6}}{1 + e^{b_0 + b_1 X_1 + \cdots b_6 X_6}}$$

The easiest approach is to do the regression plug in to get the log odds and then exponentiate and do the fraction. The child in question lives in a city so $X_1 = 1$. The pollution level is $X_2 = 35$ since pollution was given in thousands and the SES is given as $X_3 = 50$. The child was breastfed for six months so $X_4 = 6$, had parents with asthma so $X_5 = 1$ and is a boy so $X_6 = 0$. Our log odds are therefore given by

$$-.047 + -.077 + .077(35) - .082(50) + .0057(6) + 1.18(1) - 1.09(0) = -.3148$$

Don't forget the intercept! Now to get the predicted probability that the boy gets asthma we just do

$$\frac{e^{-.3148}}{1 + e^{-.3148}} = .42$$

The child has approximately a 42% chance of getting asthma.

(l) The only differences between the two children are the sex and the pollution level. We are told that the pollution level has gone down by 5 thousand particles per cm^3. For a continuous variable to get the odds ratio associated with a change of Δ we simply raise the odds ratio for a one unit increase to the power Δ. Here $\Delta = -5$ so the odds ratio for the 5 unit drop in pollution is $1.08^{-5} = 0.68$. From the printout the odds ratio for being a girl as opposed to a boy is .33. The girl should be better off on both counts. To get the overall odds ratio we multiply the individual odds ratios and get .22. The girl's odds of getting asthma are 78% lower or only about one fifth as large as her brother's!