Ordinary least squares and two-stage least squares models assume linearity in the coefficients. That is we can say that a one unit increase in X has a corresponding increase in Y. This assumption holds for linear and two-stage models because these models are only reliable with interval variables, meaning that there is a constant sequence in Y to which X can correspond.

Often times we measure things, or care about things that do not display these linear properties. These include (Yes/No) decisions, e.g. did something happen or did it not and what is the probability of it happening? These are measures for which there are not linear sequences because the value between the measures has no meaning - if something either happens or it does not then it cannot be more happening than not happening. In these cases we simply ask then, what is the probability that something will occur. To this end we use what is known as logistic regression (also probit regression works but I'll just emphasize logistic).

Discrete choice models, i.e. Logistic regression, also referred to as a "logit model," is a regression model with a categorical variable (categorical meaning exhaustive and mutually exclusive) where the variable can take a value of 0 or 1 (pass/fail; win/lose; yes/no; healthy/sick; etc.). As with other regression models we measure the relationship between the dependent variable and explanatory variables. Unlike traditional linear models in which we say a one unit increase in X has a corresponding increase in Y, we cannot interpret the coefficients of a discrete choice model as linear. If a b coefficient of X1 in a linear model has a value of 0.754 we would interpret this as meaning that for every one unit increase in X1 there is a corresponding increase in Y of 0.754. Because the space between 0 and 1 is not linear we must use the log odds of 0 or 1 occuring based on a the values taken by X1. In other words, b of X1 is a logged value of b for X1 when Y takes value 1 (yes this is confusing).

To interpret this then we must transform the log of b. There provides a number of comparisons/means of interpreting the logged coefficients:

Odds Ratio: The odds ratio represents the odds that Y will occur given a particular treatment (X1) versus the odds of it occurring in the absence of X1.

To get the odds ratio you exponentiate the value of b, e.g. exp(0.754) = 2.125 so the odds of Y occurring in the presence of X (or the mean value of X) is 2.125 to 1.

#Warning, because odds ratio relies on the direct transformation of a logged number which is unrelated to the scale of a number it treats 1:10 as the same as 10:100 and as such can yield

nonsensical results. Also the odds ratio is not at all helpful for telling you how much something changes (or is predicted to change) only that the relationship exists.

Predicted Probability: To establish the probability that something occurs means that you want to know what percent of the time something occurs given a specific value for X. Suppose you have the

following output for a model predicting whether or not people between 18-72 have health insurance (1 = yes, 0 = No):

b (se)

Intercept: 0.127 (0.034)

Age: -3.267 (1.118)

From this you can get the overall probability that sometime in this set has health insurance. This is not necessarily reflective of reality though because of policy differences, lifestyle choices and simply risk behavior with age. Suppose then that you want to know the probability of someone aged 20 having health insurance.

To assess this you would do the following:

-3.267+(20*0.127) = 0.727 #this is the logged odds that someone aged 20 has insurance. First we need to convert this to an odds ratio, so exp(0.727) = 2.068

Next we convert odds to probability, so 2.068/(1+2.068) = 67.41%, so there is a 67.41% probability that someone aged 20 has health insurance.

First Differences: Suppose now that you want to know how the probability of someone having health insurance changes as they age. We know we have a sample from age 18 to 72, so we simply take the probability of someone aged 72 having insurance and subtract it from the probability of

someone aged 18 having insurance. We can then say that as age increases from 18 to 72, the probability of their having insurance increases by 72.45%

-3.267(72*.127) = 5.877 exp(5.877) = 356.737 356.737/(1+356.737) = 99.72%

-3.267(18*.127) = -0.981 exp(-0.981) = 0.375 0.375/(1+0.375) = 27.26%

99.72-27.26=72.45 so as the age of an individual increases from 18 to 72 the probability that they have some form of health insurance increases by 72.45%. Of course knowing in the US that Medicare kicks in at 65 the high probability for someone aged 72 makes perfect sense.

Doing this in R

```
library(MASS)

library(foreign)

library(Zelig)

mydata<-read.csv(file = "~/a data file.csv")

fit1<-glm(y~x1+x2+x3, data=mydata, family=binomial(link="logit")

summary(fit)
```

calculate Odds, Probabilities and First Differences for variables of interest manually

Alternative using Zelig package in R:

```
z1<-zelig(y~x1+x2+x3, data=mydata, model="logit")

summary(z1)

x.out<-setx(z1)

s1<-sim(x.out)#give predicted probabilities for the mean value of each variable in the model

x.lo<-setx(z1, x2=min(x2)) #if you know the minmium value (you should) use that instead, same with the max below. Or you may want to know about specific ranges within x2.

x.hi<-setx(z1, x2=max(x2))

s1<-sim(z1, x=x.lo, x1=x.hi) #gives first difference as x2 moves from min to max value

summary(s1)
```