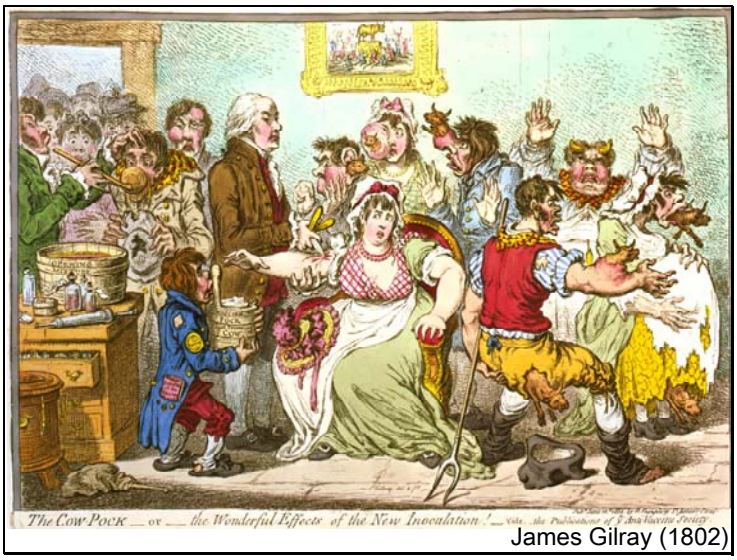
Marcello Pagano

[JOTTER 11 LOGISTIC REGRESSION]

Odds ratio, logistic function, dichotomous response, weighted least squares



James Gilray (1802)

This week we stay on our theme of regression and we are going to talk about logistic regression—we extend the idea of regression to the situation where our outcome variable is dichotomous.

To introduce the topic we look at a famous cartoon by James Gillray in 1802[1]. It deals with the reaction to what at the time was the rather revolutionary idea of vaccination. This was the first time a disease from another species was injected into man for beneficial purposes. To make the point of the anti-vaccine crowd, you see various ways of caricaturing the practice, by having cows protrude from all sorts of places.

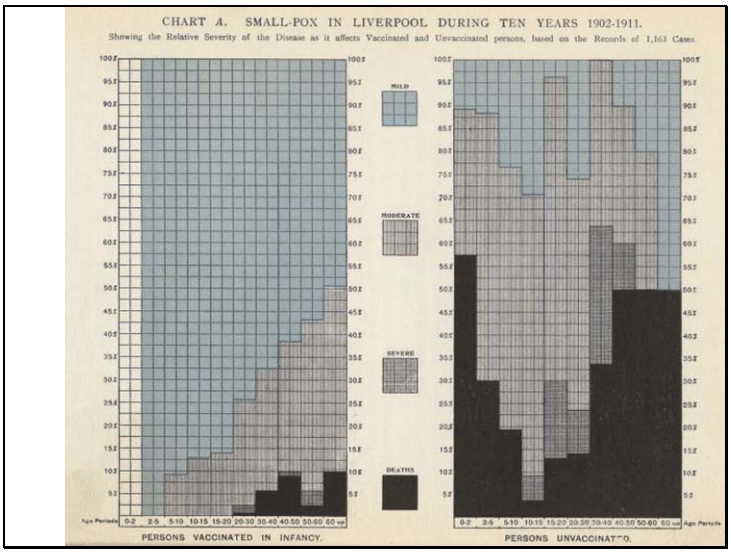
TABLE SHOWING NUMBER OF SMALL-POX CASES (948 VACCINATED AND 220 UNVACCINATED) AT EACH AGE-PERIOD, WITH RELATIVE DEGREES OF SEVERITY.

	Under 2 years		2–5		5–10		10–15		15–20		20–30		30–40		40–50		50–60		60 and up		Total	
	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.	Vacc.	Unvacc.
A. Modified discrete and discrete	No cases	3	7	3	31	6	54	9	89	1	249	11	163	No cases	63	1	20	2	10	1	686	37
B. Profuse discrete and semi-confluent	No cases	9	No cases	15	3	15	8	19	14	20	77	21	72	5	29	3	13	3	8	No cases	224	110
C. Confluent and death	No cases	17	No cases	8	No cases	5	No cases	3	No cases	9	7	10	13	9	9	6	2	5	2	1	33	73
Total	No cases	29	7	26	34	26	62	31	103	30	333	42	248	14	101	10	35	10	20	2	943	220
Deaths alone	0	17	0	8	0	5	0	1	0	4	3	6	13	8	9	5	1	5	2	1	28	60

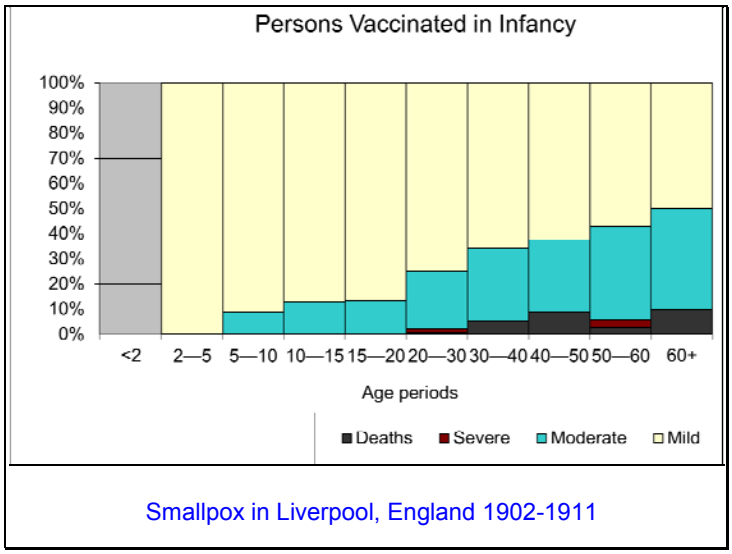
William Hanna, Studies in Smallpox and Vaccination,
William Wood and Company, New York 1913.

[1] *The Cow-Pock or the Wonderful Effects of the New Inoculation* by James Gillray (1757-1815) was published in England on June 12, 1802 by the Anti-Vaccine Society.

Smallpox was a quite lethal problem at that time, 200 years ago, as it was even 100 years ago, when this book was written[2]. It gives a thorough statistical description of a smallpox outbreak in the city of Liverpool that was part of a general outbreak they had in England in 1902-1903.



Above is a graphic from the book. It shows, as a function of age, the health impact of the epidemic. The panel on the left is for people who had been vaccinated at birth, and the panel on the right is for those not so vaccinated. One hundred years after the cartoon, vaccination had taken hold, although not everyone was vaccinated. The black bars show the number of people who died, by age. So you can see that the number of deaths is very much less for those who were vaccinated.

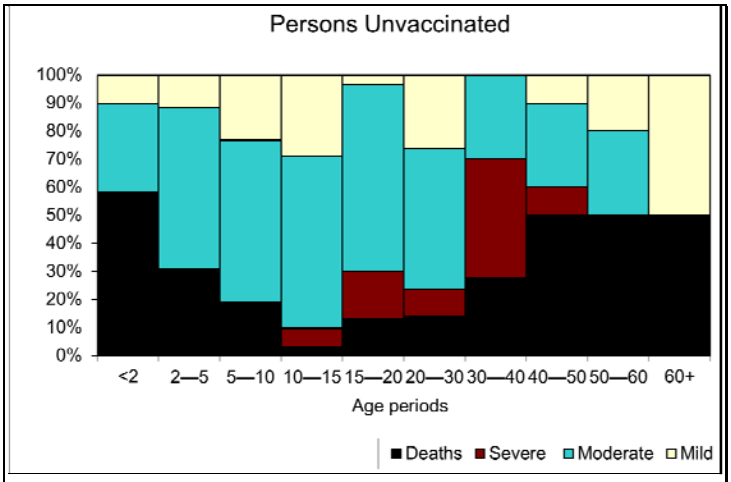


Smallpox in Liverpool, England 1902-1911

What I did was I reproduced those two graphs using modern techniques. This graph is for those who were vaccinated—the left panel above. Just focusing on the deaths we see a slight trend as people age.

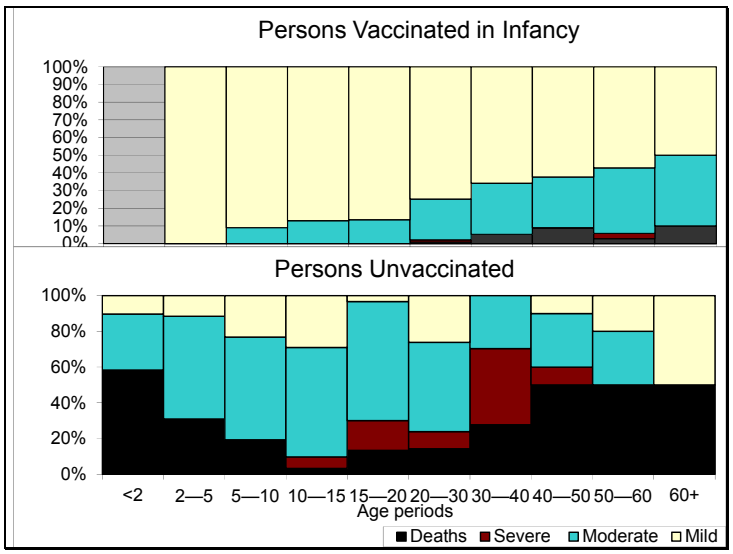
[2] William Hanna, Studies in Smallpox and Vaccination, William Wood and Company, New York 1913.

Also, if we look at the frequencies of “Moderate” effects, those too increase with age. All in all, what we see is consonant with the beliefs that the effects of vaccinations wear off with time—most of these individuals were vaccinated shortly after birth.



Smallpox in Liverpool, England 1902-1911

For those unvaccinated, we see a large number of deaths—much more black on the graph than for the previous one.



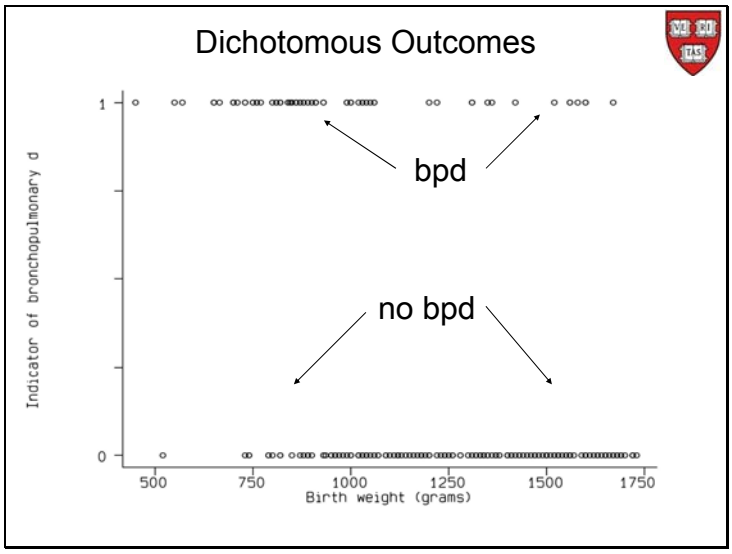
Indeed, if we place these two together, we can see that there is a big difference between the two graphs, arguing that vaccination has a great beneficial effect.

Focusing on the Vaccinated, if we look at the proportion who die in each age group we see an age effect. Certainly with the young ones, nobody below the age of 20 dies. Also, looking at the percentage who suffer Moderate, or worse, effects those too build up with age.

Smallpox in Liverpool, England 1902-1911

Age	Vaccinated			Unvaccinated		
	Numb.	Died	Mortal	Numb.	Died	Mortal
<2	0	0	0	29	17	58.0%
2—5	7	0	0	26	8	30.6%
5—10	34	0	0	26	5	19.0%
10—15	62	0	0	31	1	3.2%
15—20	103	0	0	30	4	13.3%
20—30	333	3	0.9%	42	0	14.2%
30—40	248	13	5.2%	14	8	33.3%
40—50	101	9	8.9%	10	5	50.0%
50—60	35	1	2.8%	10	5	50.0%
≥60	20	2	10.0%	2	1	50.0%
Total	943	28	2.9%	220	60	27.2%

What we are seeing is that the probability, or the proportion, who are dying is impacted by age. This reminds us of regression—we can regress the proportion who die on age. And here are the numbers to support this suggestion. If we look at the vaccinated we can see that the mortality goes up with age, almost monotonically. It is not quite smooth because of the 50 to 60 age group, but it does seem like age might affect the mortality rate, in general. Of course, we now know that the effect of vaccination does wear off with time—vaccination occurred in infancy.



Here is another example. We are looking at babies who suffer from bpd (bronchopulmonary dysplasia). This condition gets measured at about age 29 days, roughly a month after birth, amongst premature infants, and it indicates progressive lung inflammation. This study[3] was carried out on infants born weighing less than 1750 grams and each child was categorized as 0 (no bpd), or 1 (bpd) and the results are plotted above, as a function of birth age.

We can see that at the left of the graph (low birth weight) we see a higher intensity of infants with bpd, whereas on the right (relatively higher birth weight) we see a higher intensity of infants without bpd. This picture would be indicative of what we would expect to see if the probability of bpd goes down as birth weight increases. There are kids with and without bpd at either end of the scale, so we are not observing a phenomenon that deterministically decides the classification, but rather something more nuanced.

For example, if we create a window of width 250 grams, say, and slide this window from left to right, then the *proportion* of babies in the window who have bpd will go down as we move across.

[3] Van Marter, L. J., Leviton, A., Kuban, K. C. K., Pagano, M., and Allred, E. N., Maternal Glucocorticoid Therapy and Reduced Risk of Bronchopulmonary Dysplasia," *Pediatrics,* **86**, September 1990, 331-336.

Bpd by birth weight

Birthweight (gms)	Bpd	N	Prop.	Odds
0-950	49	68	0.721	2.58
951-1350	18	80	0.225	0.29
1351-1750	9	75	0.120	0.14
Total	76	223	0.341	0.52

A simple manifestation of this idea would be to take a snapshot of the window in three locations. Here I have arbitrarily chosen these three intervals, for no other reason than to make this point: the proportion with bpd goes down as we go from top to bottom. That means, as the birth weight increases. This behavior, of course, is evident in the odds of bpd, too—they go down as birth weight increases. The challenge is to be somewhat more formal in quantifying this relationship between the probability, or odds, of bpd and birth weight.

Logistic regression

In linear regression, continuous Y:

$$\mu_{Y|x} = \alpha + \beta x$$

In case of dichotomous Y:

$$\mu_{Y|x} = \Pr(Y = 1 | x) = p_{Y|x}$$

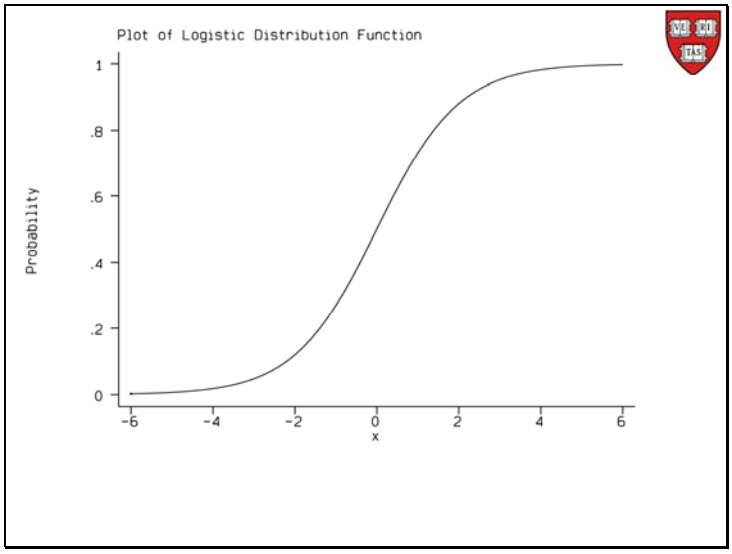
1. $p_{Y|x} = \alpha + \beta x$
2. $p_{Y|x} = e^{\alpha+\beta x}$
3. $p_{Y|x} = \dfrac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

Linear regression allowed us to quantify the relationship between a response variable and explanatory variables. For example, with simple linear regression, the regression line was the straight line relationship between the mean of the Ys and the explanatory, x.

We can mimic what we did there by first calculating the mean of the Ys. Now Y is a Bernoulli variable that takes on the values 0 and 1, so we know that its mean is the probability that it takes on the value 1. So suggestion 1, would have us fitting a straight line to that probability. That is clearly unsatisfactory because for one a straight line can be negative. So suggestion 2 would have us taking the exponential of the

straight line—that would solve the problem about becoming negative. Unfortunately, that still leaves us with the problem that this function can be greater than one and thus would not be a good model for a probability.

Suggestion 3 satisfies all the constraints on a probability. Not only that, it is monotonic in x—that means that the probability goes up with x (as x gets larger, the probability gets larger) or the probability goes down with x (as x gets larger the probability gets smaller). One can prove this mathematically, or you can convince yourself by drawing this function for yourself. You will then also see that what determines whether it goes up with x getting larger, or down with x getting larger, is the sign of the β—positive β means probability goes up with x getting larger. Note, of course, that if β=0, then the probability is not impacted by the value of x. In that case we would conclude that Y is independent of X.



If we plot this as a function of x, it looks something like this—α=0 and β=1 in this case. This is what the logistic function looks like for a positive β. If β<0, then we get the mirror image of this function that starts at one in the top-left corner and monotonically decreases to zero in the bottom right-hand corner—in other words, one minus this function.

So if we were to use this function to model the probability of bpd as a function of birth weight, we would expect a negative β. Of course, we could have switched the roles of zero and one in the definition of bpd, and then we would expect a positive β, as drawn above, because then the curve would represent the probability of *not* having bpd.

Log odds and logistic

If $$p_{Y|x} = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

then $$\ln\left(\frac{p_{Y|x}}{1-p_{Y|x}}\right) = \alpha + \beta x$$

e.g. Log odds =

$$\ln\left(\frac{p_{bpd|bwt}}{1-p_{bpd|bwt}}\right) = 4.0343 - 0.0042 \times \text{bwt}$$

where $p_{bpd|bwt}$ is the probability of bpd for a given birthweight (bwt).

If we solve the logistic function for $\alpha + \beta x$ we get the above equation for the log odds. So, simple logistic regression fits a straight line to the log of the odds.

So for example, in our bpd example, the straight line that we would fit to the log odds has intercept 4.03 and slope -0.0042. These numbers were obtained by using least squares to fit the line to the data. To be precise, it is not the least squares we are accustomed to using but, since we do not have homoscedasticity, we need to use a technique called weighted least squares. Another approach, that yields the same answer, is what is called, maximum likelihood. The technicalities associated with both are beyond the scope of this course, but this is what Stata calculates for us.

Stata:

In model form….beta

```
. logit bpd birthwt

Iteration 0:   log likelihood = -143.06998
Iteration 1:   log likelihood = -113.06062
Iteration 2:   log likelihood = -111.86443
Iteration 3:   log likelihood = -111.86031
Iteration 4:   log likelihood = -111.86031

Logistic regression                             Number of obs   =        223
                                                LR chi2(1)      =      62.42
                                                Prob > chi2     =     0.0000
Log likelihood = -111.86031                     Pseudo R2       =     0.2181

------------------------------------------------------------------------------
         bpd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     birthwt |  -.0042291   .0006409    -6.60   0.000    -.0054852   -.0029731
       _cons |   4.034291   .6957851     5.80   0.000     2.670578    5.398005
------------------------------------------------------------------------------
```

There are two Stata commands we can use. One is *logit* and it works very much like *regress*, and the output we get looks quite similar to what we get with regress, and it is interpreted much the same way.

So, in this case, we get that birth weight is significant. We see that the 95% confidence interval does not include zero, and since the coefficient is negative we conclude that, as birth weight goes up, the odds, or the probability, of getting BPD goes down. The log of the odds of getting bpd goes down by 0.0042 for every gram increase in birth weight.

Stata:

In odds form….odds ratio

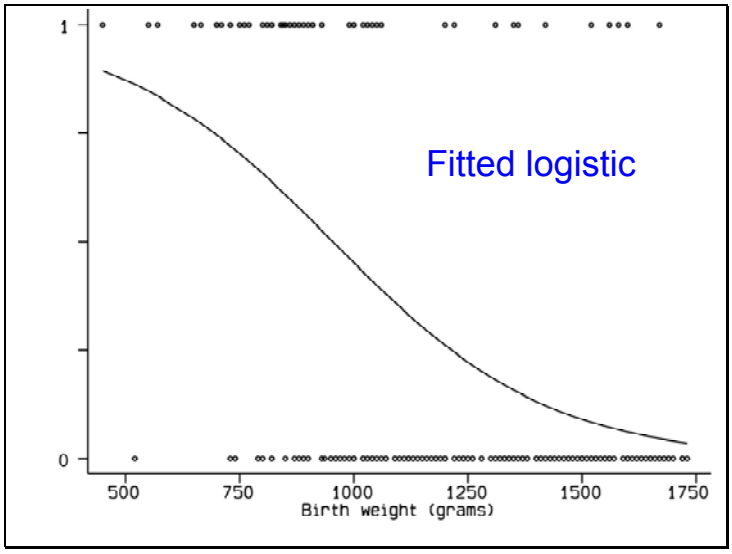
```
. logistic bpd birthwt

Logistic regression                               Number of obs   =        223
                                                  LR chi2(1)      =      62.42
                                                  Prob > chi2     =     0.0000
Log likelihood = -111.86031                       Pseudo R2       =     0.2181

------------------------------------------------------------------------------
         bpd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     birthwt |   .9957798   .0006382    -6.60   0.000     .9945298    .9970313
       _cons |   56.50287   39.31386     5.80   0.000     14.44831    220.9652
------------------------------------------------------------------------------
```

We can look at the odds directly by using the other command in Stata to fit this model, and that is the command *logistic.* It reports the odds ratio, rather than reporting the betas. Remember, we are fitting the straight line to the log of the odds, so if we report the odds that is the same as reporting the exponentials of the last screen. That means, for example, that for every gram increase in birth weight, the odds of getting bpd gets multiplied (because we are taking the exponentials) by 0.9957798. (The log(0.9957798) = -0.0042291.)

The interpretation, of course, does not change, it is just a different way to report the same results.



Here is the fitted logistic. So if we had an idealized version of the moving window we described earlier, then as it moves from left to right, this curve shows the proportion in the window with bpd.

Stata:

In model form….beta

```
. logit bpd birthwt

Iteration 0:   log likelihood = -143.06998
Iteration 1:   log likelihood = -113.06062
Iteration 2:   log likelihood = -111.86443
Iteration 3:   log likelihood = -111.86031
Iteration 4:   log likelihood = -111.86031

Logistic regression                             Number of obs   =        223
                                                LR chi2(1)      =      62.42
                                                Prob > chi2     =     0.0000
Log likelihood = -111.86031                     Pseudo R2       =     0.2181

------------------------------------------------------------------------------
         bpd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     birthwt |  -.0042291   .0006409    -6.60   0.000    -.0054852   -.0029731
       _cons |   4.034291   .6957851     5.80   0.000     2.670578    5.398005
------------------------------------------------------------------------------
```

Returning to the beta coefficients, we see that the coefficient of birth weight is small. The units of this are 1/grams, so if we were to measure the babies' birth weights in Kilograms, this coefficient would become -4.2291. One should be cognizant of the units of the measurement before opining about its size.

Observed versus fitted

Birthweight (gms)	Obsrvd bpd	Fitted bpd	N
0-950	0.721	0.646	68
951-1350	0.225	0.324	80
1351-1750	0.120	0.082	75
Total	0.341	0.341	223

One way to judge how well the model has done is to return to trichotomy we created before. In green we see the observed proportions. For each child we also now have a fitted probability of getting bpd—for a child weighing x grams, solve for p in the equation: the log(p/(1-p)) = 4.034291-0.0042291 x. The red numbers above are the averages of these p for the babies in the respective groups.

So the "fitted" values compare somewhat with the actual, observed values. There is some room for improvement.

Back to Liverpool 1902-11, vaccinated

```
. blogit dead total age, or

Logistic regression for grouped data              Number of obs   =       943
                                                  LR chi2(1)      =     19.58
                                                  Prob > chi2     =    0.0000
Log likelihood = -116.26154                       Pseudo R2       =    0.0777

------------------------------------------------------------------------------
    _outcome | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.064589   .0147107     4.53   0.000     1.036144    1.093816
       _cons |     .00361   .0020815    -9.75   0.000     .0011661    .0111764
------------------------------------------------------------------------------
```

Returning to the Liverpool outbreak, we can fit a logistic model to those data. Unfortunately, we did not have the raw data, only what we presented above, so in the reconstruction we had to cheat a little bit since we only know the ages within a category. What we did is take the midpoint of each category and used that to represent the age of all the individuals in that category. Now that we have so called blocked data—we have data only in blocks, not individual data—we can use the Stata command called *blogit,* it is the command to use for data presented in this fashion.

So here are the results for such a logistic analysis on the people who had been vaccinated prior to the outbreak in Liverpool. This says that the odds are multiplied by 1.06 every time we go up one year in age. We see that this is significant, so that age is an important explanatory variable.

Logistic growth model

Aside, population models from human ecology:

r is net population growth rate per individual

N_t is the population size at time t

$N_{t+1} - N_t = r N_t$

(Malthus, doubles every 25 years) — exponential growth

If K is number sustainable by environ

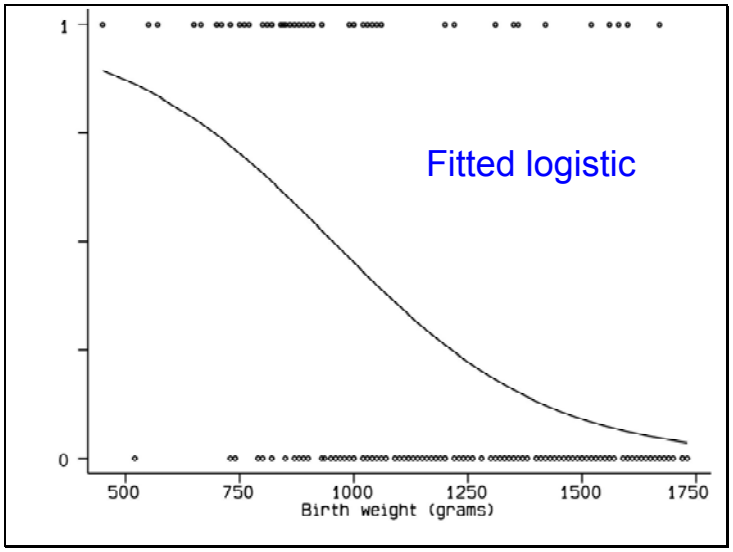
$N_{t+1} - N_t = r N_t (1 - N_t /K)$ — logistic growth

As an aside, I just wanted to point out that the logistic growth model also does appear in ecology. If you look at population size at time t, call that N_t. Suppose each individual in the population gives birth to r individuals in the next generation. Then if you look at the population growth, we see that expressed in the equation above.

This is what Malthus was concerned about almost two centuries ago, because the r he envisioned led to a doubling of the population approximately every 25 years. He was quite prescient. And that is worrisome. The value of r is all important: if r is less than 1, then we will eventually die out; r is equal to 1, then we just remain the same size; and, if r is bigger than 1, then that leads to exponential growth.

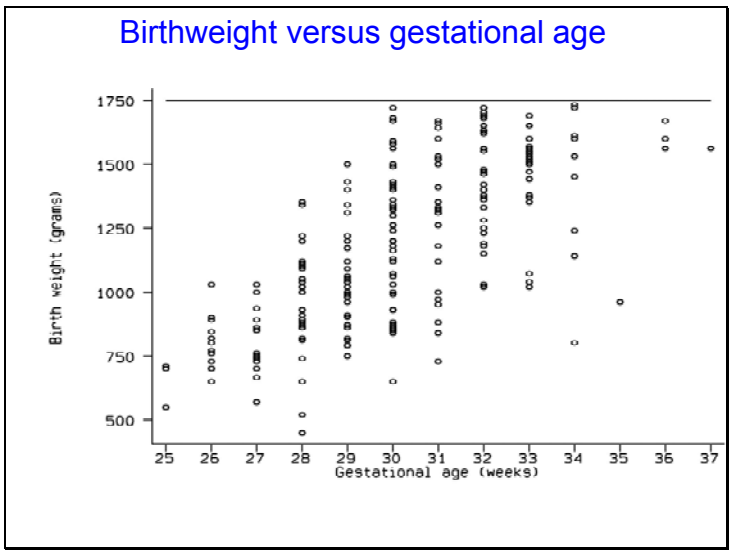
Now what happens to this model if the surroundings cannot maintain the exponential growth? For example, suppose you can only sustain K individuals in your environment. Incorporating that constraint into the growth model leads to logistic growth.

Multiple Logistic



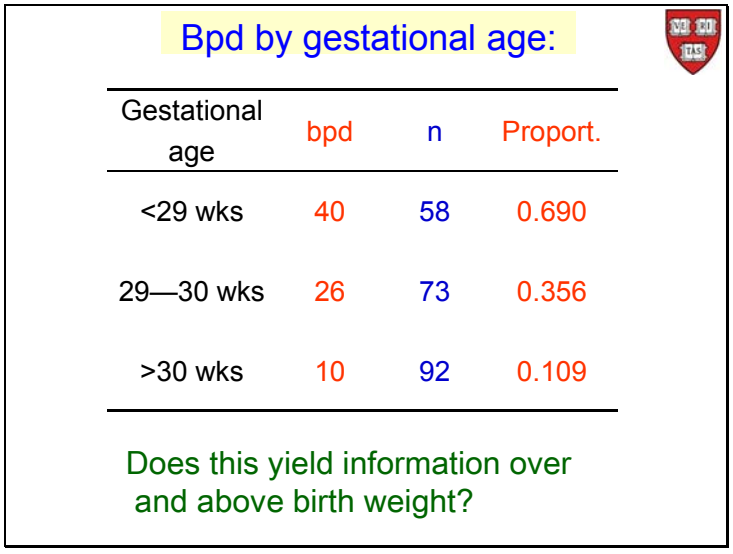
So we fitted the logistic to a single variable—birth weight—and found that the fit was good, but it left some room for improvement. Logistic regression works just like linear regression. We can do simple logistic regression with a single variable, and we can increase the number of explanatory variables to get multiple logistic regression.

We run into exactly the same problems we had with multiple linear regression: which explanatory variable will be best? Is there room for other explanatory variables? What about interaction? What about multicollinearity? And we handle them in a manner very similar to the way we did in the case of linear regression.



Consider two candidates for inclusion in the regression as explanatory variables; birth weight and gestational age. Here is a scatter plot of these two. We definitely see a relationship between the two even with the study cutoff that the babies had to weigh less than of 1750 grams at birth. This relationship is, of course, of no surprise.

So instead of birth weight, we could consider incorporating gestational age as an explanatory variable.



Once again, for exploratory purposes, let us break down the gestational age into three categories. As expected, we see a “dose response”, or monotonic relationship with proportion with bpd. A high proportion (69%) in the smallest gestational age group, have bpd. A smaller proportion (35.6%) in the middle age group, have bpd. The smallest proportion (10.9%) with bpd is in the group with the biggest gestational ages. We thus see evidence of a decrease in the chance of having bpd with gestational age.

Birthweight by gestational age

Percent with bpd:

Birth wt (gms)	Gestational age (weeks)		
	<29	29-30	>30
0-950	0.805 (n=41)	0.714 (n=21)	0.167 (n=6)
951-1350	0.412 (n=17)	0.194 (n=36)	0.148 (n=27)
1351-1750	- (n=0)	0.250 (n=16)	0.085 (n=59)

Before entering the gestational age as an explanatory variable, let us check for collinearity between gestational age and birth weight. Consider this 3-by-3 table. Once again, this is just for expository purposes, but if we had collinearity we would expect to only have single entries in a row—for example, just entries in the diagonal cells—certainly not entries in all cells but one. Indeed, in each row we see a monotonic decrease in the proportion with bpd, thus arguing that even accounting for birth weight, we see an impact of gestational age.

Having this exploratory investigation under our belts, let us fit the model with both explanatory variables.

```
. logit bpd birthwt gestage

Iteration 0:   log likelihood = -143.06998
Iteration 1:   log likelihood = -106.45551
Iteration 2:   log likelihood =  -104.6221
Iteration 3:   log likelihood =  -104.6146
Iteration 4:   log likelihood =  -104.6146

Logistic regression                             Number of obs   =        223
                                                LR chi2(2)      =      76.91
                                                Prob > chi2     =     0.0000
Log likelihood =  -104.6146                     Pseudo R2       =     0.2688

------------------------------------------------------------------------------
         bpd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     birthwt |  -.0024097   .0007925    -3.04   0.002    -.003963   -.0008564
     gestage |  -.3982616   .1129995    -3.52   0.000    -.6197366   -.1767865
       _cons |   13.02725   2.932117     4.72   0.000     8.080408    19.57409
------------------------------------------------------------------------------
```

Here is the output from Stata. As expected, we see that both coefficients are negative, so the chances of bpd go down as either birth weight or gestational age goes up. Both p values are less than 0.05, so both are considered significant. Of course, we could have arrived at the same conclusion by looking at the two confidence intervals.

```
. logistic bpd birthwt gestage

Logistic regression                               Number of obs   =        223
                                                  LR chi2(2)      =      76.91
                                                  Prob > chi2     =     0.0000
Log likelihood =  -104.6146                       Pseudo R2       =     0.2688

------------------------------------------------------------------------------
         bpd | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    birthwt |   .9975932   .0007906    -3.04   0.002     .9960448    .9991439
   gestage |   .6714864   .0758777    -3.52   0.000     .5380862    .8379587
     _cons |    1011810    2966744     4.72   0.000     3230.552    3.17e+08
------------------------------------------------------------------------------
```

We could have come to the same conclusion with the *logistic* command.

Red is observed bpd proportion, green is fitted

Birth wt (gms)	Gestational age (weeks)		
	<29	29-30	>30
0-950	0.805 0.766 (n=41)	0.714 0.514 (n=21)	0.167 0.320 (n=6)
951-1350	0.412 0.527 (n=17)	0.194 0.354 (n=36)	0.148 0.152 (n=27)
1351-1750	- - (n=0)	0.250 0.162 (n=16)	0.085 0.085 (n=59)

To investigate the fit, we can repeat the exercise we carried out above with fitted values. Averaging the fitted probabilities of each baby in each cell we get the green numbers above. This shows an improvement over the model with only birth weight as an explanatory variable.

Indicator Variables

toxemia = 1 if mother had toxemia
= 0 otherwise

```
. logit bpd toxemia

Iteration 0:   log likelihood = -143.06998
Iteration 1:   log likelihood = -141.64339
Iteration 2:   log likelihood = -141.63956
Iteration 3:   log likelihood = -141.63956

Logistic regression                               Number of obs   =        223
                                                  LR chi2(1)      =       2.86
                                                  Prob > chi2     =     0.0908
Log likelihood = -141.63956                       Pseudo R2       =     0.0100

------------------------------------------------------------------------------
         bpd |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     toxemia |   -.7719484   .4821775    -1.60   0.109    -1.716999    .1731021
       _cons |   -.5717863   .1494999    -3.82   0.000    -.8648008   -.2787718
------------------------------------------------------------------------------
```

$$\ln\left(\frac{\hat{p}_{Y|\text{toxemia}}}{1-\hat{p}_{Y|\text{toxemia}}}\right) = -0.572 - .772\ \text{toxemia}$$

Let us look at indicator variables with logistic regression because it should remind us of something we have already seen. Recall that we looked at toxemia as an indicator variable that indicated whether the mother suffered from toxemia during the pregnancy. Here is the result of fitting a logistic regression with just toxemia as the explanatory variable.

We see that the coefficient is -0.77 and is judged not to be significant. But here is the best fitting model if we include toxemia.

$$\ln\left(\frac{\hat{p}_{Y|\text{toxemia}}}{1-\hat{p}_{Y|\text{toxemia}}}\right) = -0.572 - .772\ \text{toxemia}$$

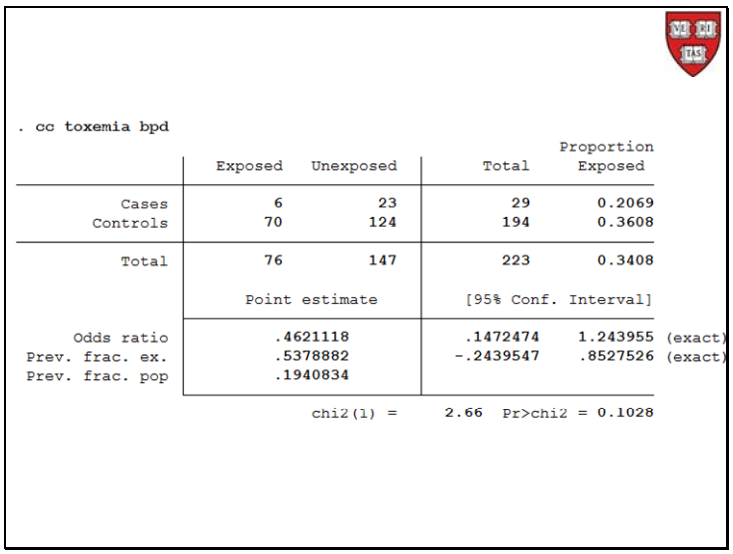
$$\ln\left(\frac{\hat{p}_{Y|\text{has toxemia}}}{1-\hat{p}_{Y|\text{has toxemia}}}\right) = -0.572 - .772$$

$$\ln\left(\frac{\hat{p}_{Y|\text{no toxemia}}}{1-\hat{p}_{Y|\text{no toxemia}}}\right) = -0.572$$

$$\ln(\text{OR}) = -.772$$

$$\Rightarrow \text{OR} = e^{-.772} = 0.46$$

So the log odds of bpd are as above. The difference between the log odds when the mother has toxemia and when the mother does not have toxemia, is the log of the odds ratio, and that is -0.772. So the odds ratio is 0.46.

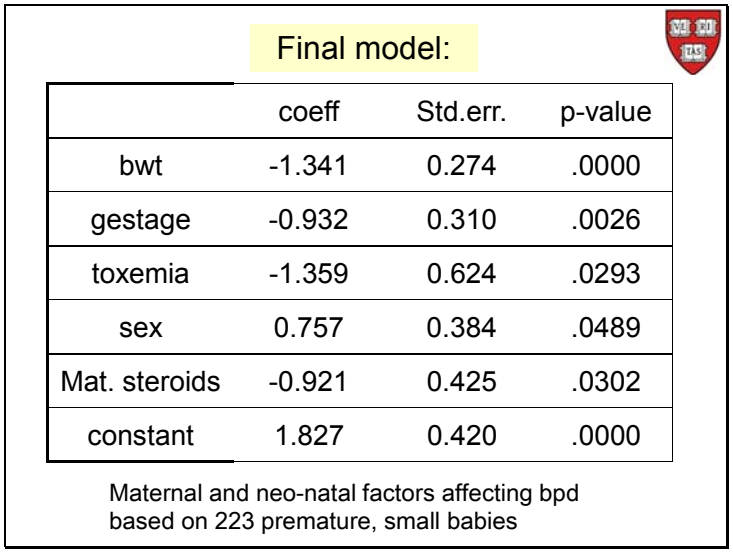


```
. cc toxemia bpd

                                                      Proportion
                |   Exposed   Unexposed  |      Total   Exposed
----------------+------------------------+------------------------
          Cases |           6          23  |         29      0.2069
       Controls |          70         124  |        194      0.3608
----------------+------------------------+------------------------
          Total |          76         147  |        223      0.3408
                |                        |
                |      Point estimate    |    [95% Conf. Interval]
----------------+------------------------+------------------------
     Odds ratio |          .4621118      |     .1472474    1.243955 (exact)
Prev. frac. ex. |          .5378882      |    -.2439547    .8527526 (exact)
Prev. frac. pop |          .1940834      |
----------------+------------------------+------------------------
                                  chi2(1) =     2.66  Pr>chi2 = 0.1028
```

Alternatively, we can consider the 2x2 table of bpd versus toxemia. When we analyze that table, as above, we see that the odds ratio is the same 0.462 as we got from the logistic regression analysis.

So an indicator variable and logistic regression is just another way of looking at the tables we had analyzed before. What the logistic regression approach allows us to do now is to analyze tables considering a collection of dimensions—any number of explanatory variables—together with interaction terms, etcetera.



Final model:

	coeff	Std.err.	p-value
bwt	-1.341	0.274	.0000
gestage	-0.932	0.310	.0026
toxemia	-1.359	0.624	.0293
sex	0.757	0.384	.0489
Mat. steroids	-0.921	0.425	.0302
constant	1.827	0.420	.0000

Maternal and neo-natal factors affecting bpd
based on 223 premature, small babies

Indeed, the final model for these data is above, and we see a collection of dichotomous (toxemia, sex, maternal steroids (yes/no)) and continuous explanatory variables (birth weight and gestational age). This is a much richer class of models to fit to a dichotomous response variable than is afforded by analysis as

2x2 tables. Note that the final model includes toxemia—possibly because toxemia happens later in the pregnancy and is related to gestational age.