

Logistic Multiple Regression

STAT 102

- **Logistic Regression** is suitable for situations in which the Y variable is categorical rather than continuous.
- We will study only the most common setting, in which the Y variable is binary (*ie*, has only two possible values).
- Lecture 21 discussed the situation having one continuous X -variable as a possible predictor.
- We now turn to some examples involving more than one X -variable.
- The X -variables may be either continuous or categorical. **BUT**
- Settings in which **all** X -variables are categorical have some special features, and will be discussed in Lecture 23.

Two Examples

- This Lecture analyzes two separate examples.
- If we do not have time in lecture to completely cover both; you should still read and understand both.

Example 1: Prediction of the Risk of Coronary Heart Disease.

Example 2: March Madness.

- How well do the seedings predict who will win?
- Is there a difference in this regard between the Men's Tournament and the Women's?

Example 1:

Prediction of the Risk of Coronary Heart Disease

DATA: The data for this example is taken from the Framingham Heart Study.

- This study was an early, large public health study designed to investigate risk factors for **Coronary Heart Disease** [“CHD”], and to study the course of that disease, when it occurs.
- Our data consists of health variables measured for a population of 1408 adult health professionals, age 45-62.
- These variables were measured at the start of the study (in the '50s).
- At that time these subjects were also given an intensive (and expensive) series of examinations of coronary health status.
- Other tests were given, and the patients were followed-up at regular intervals over their lifetime, but we will look only at information from the initial intake exam.

Variables in our Data

All variables measured at initial exam

- **Heart Disease?** [This is our Y -variable]: A 0-1 indicator as to whether the initial intensive exam found evidence of CHD.
- **Age:**
- **SBP:** Systolic blood Pressure
- **DBP:** Diastolic Blood Pressure
- **CHOL:** Cholesterol level
- **FRW:** A measure of weight, adjusted for patient's height and age
- **CIG:** Number of cigarettes smoked/day. [Self Report!]
- **Sex:**

Logistic Regression Model

- Goal is to estimate $p(\mathbf{x}) = \Pr(Y = \mathbf{0}|\mathbf{x})$.
- Here, \mathbf{x} represents the independent variables in the model.
- The assumed form of $p(\mathbf{x})$ is as follows:

$$p(\mathbf{x}) = \ell(\theta) = \frac{e^\theta}{1 + e^\theta} \text{ where}$$

- Here, θ is a general symbol used to denote the value of a linear function of the independent variables.
- Thus, if the model has m quantitative factors and one qualitative factor (Sex), the model for θ is

$$\theta = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m + \beta_{[\text{Sex}]}$$

- **Be careful.** JMP defines p to correspond to the value $Y = \mathbf{0}$. Most other software defines it to correspond to $Y = 1$.

Analysis

- JMP performs the needed analyses.
- It provides estimates b_0, \dots, b_m of the (logistic) regression intercept and slope coefficients β_0, \dots, β_m ,
- and estimate(s) $b_{[\text{Sex}]}$ of the (logistic) ANOVA coefficient(s) $\beta_{[\text{Sex}]}$.
- JMP also produces tables needed for hypothesis tests.
- The estimates produced by JMP are called **Maximum Likelihood Estimates**.
- Maximum Likelihood is a general method for producing desirable estimates in many statistical settings.

Basic Output

from a model with all available factors

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	63.094	7	126.1882	<.0001
Full	671.572			
Reduced	734.666			
RSquare (U)		0.0859		
Observations		1393		

- In the whole model test table **ChiSquare = $2 \times (-\text{LogLikelihood})$**
- For this, use the Chi-squared table with 7 df
- This tests H_0 : **All** regression slopes & factor effects = **0**.
- $n = 1406$, but **Observations = 1393** because 13 patients have some missing data and are automatically excluded from the JMP analysis
- We will exclude all these patients from future analyses.
- Like in usual multiple regression, $\text{RSquare (U)} = 63.094 / 734.666$.

Effect Likelihood Ratio Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
AGE	1	17.613	<.0001
SBP	1	14.684	0.0001
DBP	1	0.142	0.7060
CHOL	1	8.733	0.0031
FRW	1	2.023	0.1549
CIG	1	4.042	0.0444
SEX	1	34.048	<.0001

- DF here is like in usual tables: 1 DF for each regression coefficient and 1 DF for Sex because there are two sex categories.
- The **L-R ChiSquare** values here come from a “Full” – “Reduced” analysis using the **Whole Model Test** table like in usual tables.
- BUT there’s a secret! We’ll give an example later.
- The **Prob>ChiSq** entries come from these statistics and the Chi-squared table with DF Degrees of Freedom.

Parameter Estimates

Term	Estimate	Std Error	ChiSq	Pr>ChiSq	Est'te at
Intercept	8.8817	1.0237	75.28	<.0001	
AGE	-0.0625	0.015	17.37	<.0001	45
SBP	-0.0148	0.00389	14.58	0.0001	100
DBP	-0.00288	0.00762	0.14	0.7059	80
CHOL	-0.00446	0.00151	8.78	0.0031	180
FRW	-0.0058	0.00406	2.04	0.1530	110
CIG	-0.01231	0.00609	4.09	0.0432	5
SEX[FEM]	0.45305	0.07882	33.04	<.0001	Fem

For log odds of 0/1

- The **Estimates** are the important part of this table.
- The ChiSquare values and **Std Errors** here are from a different formula than that used in the **Effect LRT** table.
- Consequently, the P-values here may not exactly equal those in the preceding table. But they should be close.
- We've added a **column of variables describing a person**. The est for this person is $\theta = 8.8817 - .0625 \times \mathbf{45} - .0148 \times \mathbf{100} - \dots + \mathbf{.45305} = 3.307$ &

$$p = e^{\theta} / (1 + e^{\theta}) = \mathbf{.965}$$

Re: The LR Chi Sq entries in the Effect Tests table

- LR Chi Sq values are $2 \times (\text{Full} - \text{Reduced})$.
- Here's the Whole Model table for an analysis **without** DBP:

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	63.023	6	126.0459	<.0001
Full	671.643			
Reduced	734.666			

- Hence the LR ChiSq for testing DBP after controlling for all other variables is

$$2(671.643 - 671.572) = 2 \times .072 = .144$$

This is the value in the earlier Effect Tests table (except for round-off).

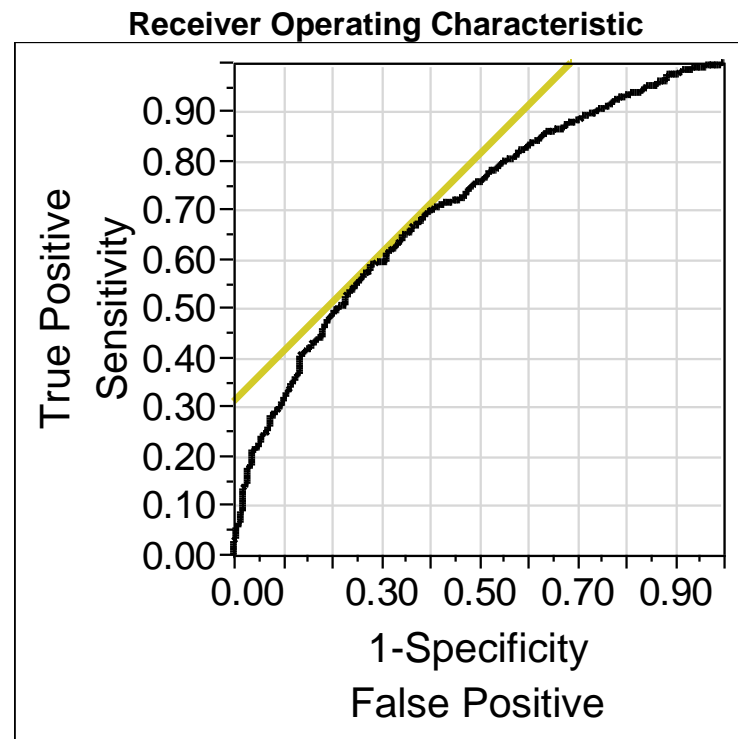
Estimates in JMP

- If you go to “Save Probability Formulas” under the red arrow, you get columns headed

Lin[0]	Prob[1]	Prob[0]	Most likely Heart Disease?
3.3073	0.03532	0.9647	0

- The entries in **green** are those in the row for our special **Person**
- The Column **Lin[0]** gives the values of θ .
- “Most Likely” just tells whether $\text{Prob}[1] > 1/2$.

You can get the **ROC Curve**



Using Heart Disease?='0' to be the positive level
Area Under Curve = 0.7036

ROC Table

Here are a few lines of the Table

Prob	1-Specificity	Sensitiv	Sens+Spec		True Pos	True Neg	False Pos	False Neg
0.2028	0.4098	0.7166	0.3069		220	641	445	87
0.2026	0.4107	0.7166	0.3060		220	640	446	87
0.2023	0.4116	0.7231	0.3116	*	222	639	447	85
0.2018	0.4134	0.7231	0.3097		222	637	449	85
0.2016	0.4153	0.7231	0.3079		222	635	451	85

- The choice $p_0 = .2023$ gives the rule that maximizes **Sens+Spec**
- This rule is **Declare 0** (=NoCHD) if $\hat{p} \geq p_0 = .2023$. Otherwise, Declare **1** (=CHD).
- IF half the population at large has noCHD and half does then we can estimate that in future this declaration will be correct with

$$P(\text{Corr}) = \frac{1}{2} P(\text{Corr}|\text{noCHD}) + \frac{1}{2} P(\text{Corr}|\text{CHD}) \approx \frac{1}{2} .5884 + \frac{1}{2} .7231 = .6558.$$

However

- The data is a random sample of health care workers. Hence
- **The best estimate** for the proportion of the entire population of health care workers (ages 45-62) with no CHD is

$$P(\text{noCHD}) \approx \frac{\#\{\mathbf{0} \text{ in sample}\}}{\text{sample size}} = \frac{1086}{1393} = .7796$$

- We can now estimate for the preceding classification rule

$$P(\text{Corr}) \approx .7796 \times .5884 + .2204 \times .7231 = .6181.$$

- Working out the arithmetic a little more carefully yields

$$P(\text{Corr}) \approx \frac{1086}{1393} \times \frac{639}{639 + 447} + \frac{307}{1393} \times \frac{222}{222 + 85} = 639 + 222 = \text{Total "True"}.$$

We See That

- **The best estimate for the Prob of a correct classification in a population like that at hand is**

Total True in the ROC Table

The best estimate for the Prob of a correct classification:

Here are the relevant rows of our ROC table

Prob	1-Spec	Sens	Sens+Spec 1	True Pos	True Neg	False Pos	False Neg	Total True
0.5122	0.0147	0.0782	0.0634	24	1070	16	283	1094
0.5120	0.0147	0.0814	0.0667	25	1070	16	282	1095
0.5052	0.0147	0.0847	0.0700	26	1070	16	281	1096
0.5045	0.0157	0.0847	0.0690	26	1069	17	281	1095
0.5043	0.0166	0.0847	0.0681	26	1068	18	281	1094
0.5040	0.0166	0.0879	0.0714	27	1068	18	280	1095

Hence the above rule, with $p_0 = .5052$, yields an estimated probability of true classification of $P(\text{Corr}) \approx \frac{1096}{1393} = .7868$ vs

$P(\text{Corr}) \approx \frac{861}{1393} = .6181$ for the rule in the Table with a *.

Important Note:

- The preceding choice of rule, with $p_0 = .5052$, has assumed that False Positives and False Negatives are equally serious.
- That's not true here.
- As an example: Let's assume
- Diagnosing a nonCHD person as having **CHD** (a “False Pos”) causes one “unit” of pain and discomfort.; and
- Diagnosing a CHD person as being **nonCHD** (a “False Neg”) causes **10** “units” of pain and discomfort.
- Then we want to minimize $1 \times \#FalsePos + 10 \times \#FalseNeg$
- This occurs at $p_0 = .1033$ with Sens = .9642 & Spec + .2072.
- Then have many FalsePos (861) but very few FalseNeg (11)

Example 2

March madness

Every March the NCAA holds its annual Div 1 College Basketball Tournaments (for Women and for Men).

Teams are seeded into four “brackets” of 16 in each tournament. In each bracket the teams are “seeded” from 1 (best) to 16 (worst). They play a single elimination tournament.

At the end, the winners of the four brackets hold a four team single-elimination mini-tourney.

How good are the seedings? (How often does the better seeded team win?)

For the past few years it has been claimed that the women are more accurately seeded than the men. (If so, this may be because the ability differences among the women’s teams are wider.)

Is this the case?

Data: The tournament results for both men and women for 2003 – 2006 have been coded. [We haven't gotten around to entering 2007 yet!]

Each row of data corresponds to one game

The columns are:

- **DIFF** = The difference in seeding between the higher and lower seeded teams. These values are ≤ 0 .
- **WIN** = 1 if the Favorite team (higher seeded) wins and 0 if that team loses. (In the final two rounds it is possible to have $\text{DIFF} = 0$. If so, one team has (arbitrarily) been listed first, and $\text{WIN} = 1$ if that team wins.)
- **ROUND** = round of the tournament
- **Year** and **M or W** are self explanatory
- **PCTHIGH** and **PCTLOW** are the regular season winning proportions for the Higher and Lower seeded teams, respectively.

- We ran a logistic multiple regression with
 - $Y = \text{WIN}$, and
- Model Effects:
 - **DIFF**
 - **M or W**
 - **Year**
 - **ROUND**

- **Year** and **ROUND** are included primarily as control variables, but strong significance of either one could be of interpretive interest
- **DIFF** is expected to be the dominant effect.

Because we have coded $\text{DIFF} \leq 0$, and because JMP estimates $\text{Prob}(Y=0)$ – *ie*, the probability of an upset – we expect the DIFF coefficient to be >0

- **M or W** is the effect of primary interest.
If the coefficient for Men is significant & >0 that will show that underdogs do better in the Men's tourney, even after taking account of the seeding (=DIFF) and the Year and Round.

Analysis

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	54.849	11	109.70	<.0001
Full	238.033			
Reduced	292.882			

RSquare (U) 0.1873
Observations 504
For log odds of 0/1

Effect Likelihood Ratio Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
DIFF	1	71.839	<.0001
Year	4	3.398	0.4936
M or W	1	3.009	0.0828
ROUND	5	10.499	0.0623

These tables look much as expected AND

M or W is **NOT STATISTICALLY SIGNIFICANT** at $\alpha = 0.05$

Parameter Estimates

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.709	0.255	7.756	0.0054
DIFF	0.295	0.040	53.777	<.0001
Year[03]	-0.071	0.214	0.109	0.7410
Year[04]	0.058	0.212	0.076	0.7831
Year[05]	0.137	0.210	0.426	0.5139
Year[06]	-0.517	0.334	2.403	0.1211
M or W[M]	0.225	0.130	2.980	0.0843
ROUND[1]	-0.343	0.275	1.554	0.2125
ROUND[2]	0.052	0.286	0.033	0.8550
ROUND[3]	-0.840	0.329	6.506	0.0108
ROUND[4]	-0.255	0.376	0.461	0.4973
ROUND[5]	-0.125	0.478	0.069	0.7934

These also look much as expected: Especially, $\text{DIFF} > 0$

Overall, ROUND is not quite significant at 0.05, but the prominent feature is that $\text{ROUND}[3] < 0$. This *suggests* that *perhaps* favorites do their best in round 3.

Is the Seeding Committee Doing the Best it Can?

An interesting feature appears when DIFF in PCT is included in the model.

Here is the output:

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	58.658	12	117.32	<.0001
Full	234.224			
Reduced	292.882			

RSquare (U)	0.2003
Observations	504

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
DIFF	1	1	52.872	<.0001
Year	4	4	3.073	0.5457
M or W	1	1	3.196	0.0738
Diff in PCT	1	1	7.618	0.0058
ROUND	5	5	11.013	0.0511

- Here, Diff in PCT is also statistically significant (in addition to DIFF). It has a noticeable effect on U, so it is also numerically relevant.
- The coefficient of Diff in PCT is < 0 , as one would expect. (You can check this in the Parameter Estimates table.)

This indicates that even after controlling for seeding, teams with higher regular season winning percentages are less likely to be upset by lower seeded teams (and vice-versa).
- This suggests that (if it wanted to do so) the seeding committee could produce more accurate seedings by paying more attention to the winning percentage of the teams.