Logistic Regression :
Regression for Binomial Responses

Agresti : Categorical Data Analysis, Second Edition, Wiley and Sons, 2002
*Also, take **Categorical Data Analysis** at EPH*

What is does : Generalized Linear Models for binary (0, 1) responses.

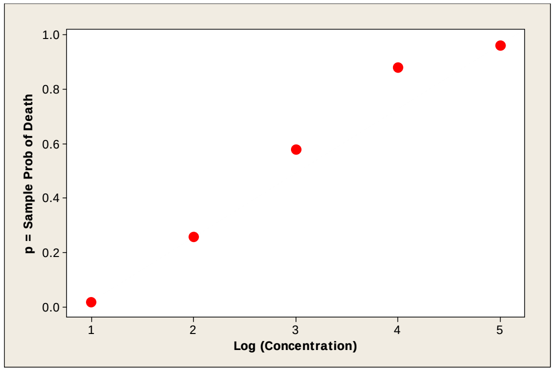
Cases where it's used :

- Model the probability of getting lung cancer
- Model the probability that someone believes in UFO's
- Model the probability of favoring the death penalty

Can be used on Categorical Data (table data) where one variable has two outcomes.

***Example** : **Ticks and Fungus.** Michael Benjamin (et all, J. of Medical Entomology, 2002), found the following mortality rates when the fungus* Metarhizium anisopliae *was sprayed on deer ticks in at five concentration levels :*

Concentration (log scale)	***Number of Insects***	***Number Dead***	***Sample Proportion Dead*** $\hat{p}$
1	*50*	*1*	*0.02*
2	*50*	*13*	*0.26*
3	*50*	*29*	*0.58*
4	*50*	*44*	*0.88*
5	*50*	*48*	*0.96*



How to Model the probability of Success?

Could try the usual regression model :

$$p = \beta_0 + \beta_1 X + \varepsilon$$

$\varepsilon \sim N(0, \sigma)$ for **<u>all</u>** p

where X is the log(concentration).

Problems :

- If $\beta_1 \neq 0$, large values of X will predict probabilities p >1 or p<1, which is impossible *(think about log concentrations of 0 or 6).*
- We would expect that as the concentration of fungus increases, at some point p should stay at about 1 *(i.e. once at a concentration where all ticks are dead, doubling the concentration means they're still dead!)* Similary, for low concentrations, p should be about zero.
- ***Variance of binomial** is* $np(1-p)$: obviously, **<u>not</u>** constant variance over the range of values for p

SO : usual regression model is not appropriate!!!

Let's consider only two concentrations, say Low and High

Concentration (log scale)	***Number of Insects***	***Number Dead***	***Sample Proportion Dead*** $\hat{p}$
2	*50*	*13*	*0.26*
5	*50*	*48*	*0.96*

Now, think about **conditional probabilities**:

$$\Pr(Dead \mid Low) = \frac{13}{50} = 0.26$$

$$\Pr(Dead \mid High) = \frac{48}{50} = 0.96$$

Let's say we're betting on ticks *: what are the odds a tick dies given it receives low concentration?*

$$Odds(Dead \mid Low) = \frac{13}{37} = 0.35$$

i.e. odds are 3 to 1 against dying

$$Odds(Dead \mid High) = \frac{48}{2} = 24$$

i.e. odds are 24 to 1 in favor of dying

NOW : Note that we can also compute the odds from the conditional probabilities :

$$Odds(Dead \mid Low) = \frac{\Pr(Dead \mid Low)}{1 - \Pr(Dead \mid Low)} = \frac{0.26}{1 - 0.26} = 0.35$$

$$Odds(Dead \mid High) = \frac{\Pr(Dead \mid High)}{1 - \Pr(Dead \mid High)} = \frac{0.96}{1 - 0.96} = 24$$

Taking logs, we get the *log odds* ***, known as the*** *logit transformation*

$$\ln[Odds(Dead \mid Low)] = \ln(0.35) = -1.1$$

$$\ln[Odds(Dead \mid High)] = \ln(24) = 3.1$$

Define X =1 if the concentration is high, 0 is concentration is low.

$$\ln[Odds(Dead \mid X)] = -1.1 + 4.2X$$

In terms of probabilities,

$$\ln\left(\frac{p}{1-p}\right) = -1.1 + 4.2X$$

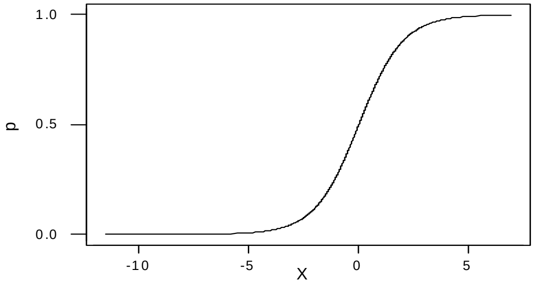
(This looks suspiciously like regression)**!!**
Now :

- Allow X to take on the original five log(conc) values.
- Clearly, the probability of death increases with increasing log(concentration)
- We can model the log odds probability of dying as a **linear** function of concentration
- Naturally, we won't achieve a perfect fit – there will be some error.

We use the LOGISTIC REGRESSION MODEL :

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X + \varepsilon$$

$$\varepsilon \sim N(0,\sigma^2)$$



Notes on Logistic Function :

- Note that p and X are **NOT** linearly related. The relationship between p and X is given by (just exponentiate both sides and solve for p)

$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Note that for large values of X, p approaches 1; for small values of X, p approaches zero. This is what we wanted!

- However, $\ln\left(\frac{p}{1-p}\right)$ and X **ARE** linearly related.

- Logistic regression accounts for the fact the some groups may have more observations than others in calculating regression results (called **weighted regression**). *Not relevant for tick data since have 50 observations at each concentration level.*

- Logistic regression may also consider categorical factors *(examples of this in a bit)*

Logistic Regression In MINITAB, use `Stat` → `Regression` → `Binary Logistic Regression`. You can enter raw data (i.e. with a response that is 0 or 1), or you can enter summarized data in three forms. For the tick data, use the successes / trials option (we'll assume a dead tick is a success).

Logistic Regression In SPSS, use `Analyze` → `Regression` → `Binary Logistic Regression`. Enter continuous predictors under `Covariates`. Click on `Categorical` to enter categorical predictors.

Let's look at MINITAB output section by section:

First is a summary of total successes and failures in all groups :

```
Response Information

Variable  Value    Count
Died      Success    135
          Failure    115
TOTAL     Total      250
```

Next is a test of significance for each parameter :

```
Logistic Regression Table

                                                     Odds     95% CI
Predictor              Coef   SE Coef       Z      P  Ratio  Lower  Upper
Constant           -4.52037  0.562198   -8.04  0.000
log(concentration)  1.61817  0.188381    8.59  0.000   5.04   3.49   7.30
```

Our Model :

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1\log(Conc)+\varepsilon$$

*log odds prob of dying = constant + slope*log(concentration) + errors*

Hypothesis Test :

$$H_0:\beta_1=0$$ *(death rate is unaffected by concentration)*

$$H_a:\beta_1\neq 0$$ *(death rate is affected by concentration)*

This is tested by calculating a z-statistic akin to that in simple linear regression :

$$z=\frac{b_1}{SE(b_1)}=\frac{1.62}{0.19}=8.6$$

Under the null hypothesis, z has an approximately standard normal distribution. P-value=Prob(8.6 or larger in a standard normal distribution)= 0.000.

Reject null hypothesis and conclude that concentration is a significant predictor of death rates of ticks.

Interpreting Logistic Regression Coefficients

- The Logit transformation is a **monotonic increasing function** – that is, positive regression coefficients mean that higher values of covariates are associated with increased probability of event (just like regular regression)

Example : Tick Data** : Slope = 1.6 That is : **higher concentrations are associated with increased probability of tick death

HOWEVER : **THE USUAL INTERPRETATION OF REGRESSION COEFICIENTS DOESN'T WORK!**

What you want to think – if we double the concentration, we double the probability a tick will die. ***NOT TRUE!***

Use Brain : *at very low concentrations, doubling the concentration will have little effect on tick death. For some middle concentrations, double the concentration will kill many more ticks. At high concentrations, doubling the concentration has almost no additional effect (since the ticks are already all dead!)*

SO : How do we numerically interpret the magnitude of regression coefficients relative to probabilities of tick death?

Our Model :
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$$

Exponentiate, solve for p :

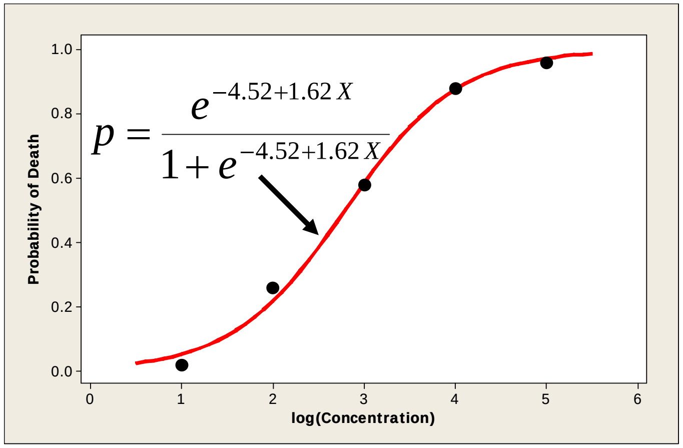
$$p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}} \quad \text{or for our data} \quad p = \frac{e^{-4.52+1.62\,X}}{1+e^{-4.52+1.62\,X}}$$

This formula can be used to get the estimated probability for each value of X

Concentration (log scale)	*Sample Proportion Dead* $\hat{p}$	*Predicted Proportion dead :* $p = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$
-1	-	*0.002*
0	-	*0.01*
1	***0.02***	***0.05***
2	***0.26***	***0.22***
3	***0.58***	***0.58***
4	***0.88***	***0.88***
5	***0.96***	***0.97***
6	-	*0.995*
7	-	*0.999*

The plot below shows the calculated regression curve and the sample observed probabilities. The estimated curve is

$$\log\left(\frac{p}{1-p}\right) = -4.52 + 1.62\log(Concentration)$$



Odds Ratio

MINITAB estimates the odds ratio for **a one-unit increase in the explanatory variable** (i.e. log concentration) : *that is, for each one-unit increase in log-concentration, the odds ratio is estimated to increase by 5.04.*

```
                                                         Odds     95% CI
Predictor                 Coef   SE Coef      Z      P  Ratio  Lower  Upper
Constant              -4.52037  0.562198  -8.04  0.000
log(concentration)     1.61817  0.188381   8.59  0.000   5.04   3.49   7.30
```

The odds ratio is determined simply by exponentiating the predicted slope : $e^{1.61} = 5.04$

This is because :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \quad \text{so} \quad Odds = \frac{p}{1-p} = e^{\beta_0 + \beta_1 X}$$

When X = 0 : $Odds = \dfrac{p}{1-p} = e^{\beta_0}$

When X = 1 : $Odds = \dfrac{p}{1-p} = e^{\beta_0 + \beta_1}$

Odds Ratio = Ratio of the odds : $\dfrac{e^{\beta_0 + \beta_1}}{e^{\beta_0}} = e^{\beta_0 + \beta_1 - \beta_0} = e^{\beta_1}$

and this is true for any one unit change in X.

Conc. (log scale)	Sample Proportion Dead $\hat{p}$	Sample Odds	Sample Odds Ratios, successive categories	Predicted Proportion Dead	Predicted Odds	Predicted Odds Ratios, successive categories
1	0.02	0.02		0.05	0.055	
2	0.26	0.35	17.5	0.22	0.277	5.04
3	0.58	1.38	3.94	0.58	1.397	5.04
4	0.88	7.33	5.31	0.88	7.045	5.04
5	0.96	24	3.27	0.97	35.534	5.04

The confidence interval for the odds ratio is calculated as

$$e^{\beta_1 \pm 1.96 StDev(\beta_1)}$$

Next part of MINITAB output :

```
Log-Likelihood = -95.384
Test that all slopes are zero:
    G = 154.204, DF = 1, P-Value = 0.000
```

What is this??!?!?

LIKELIHOOD

- The likelihood of a model is **the probability of observing our sample data under that model.**

Example : *toss a coin 4 times, get four heads.*

- *The Likelihood of the data under the model of a fair coin is $(1/2)^4$=1/16.*
- *The Likelihood of the data under the model of a two-headed coin is 1^4= 1.*
- *The two-headed coin model is more 'likely', in fact it is the <u>maximum likelihood estimator.</u>*

LOG - LIKELIHOOD = Natural Log of the Likelihood
(what a surprise!)

Consider the following Hypothesis Test:

H_0 : All regression coefficients are zero
H_a : At least one regression coefficient is not zero

For this data, only one regression coefficient :

$$H_0 : \beta_{Concentration} = 0 \qquad H_a : \beta_{Concentration} \neq 0$$

This is equivalent to comparing two models :

Model Under Null Hypothesis (no slope) :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \varepsilon$$

Model Under Alternative Hypothesis :

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X + \varepsilon$$

Now : *we saw 135 dead ticks out of 250 total ticks, i.e. a sample probability of dying of 0.54.*

If the Null hypothesis is true, this should be the probability of a tick dying at ALL concentration levels!

SO : Likelihood of data under our null hypothesis model =

$$.54^{135}.46^{115}$$

The log of this number (the ***Log Likelihood*** *) is -172.49.*

Now : *in model with slope, we estimate a* ***different*** *probability of death at each concentration level (now we think concentration has an effect on tick death) :*

Log (Conc)	*1*	*2*	*3*	*4*	*5*
Estimated Probability $p = \frac{e^{-4.52+1.62X}}{1+e^{-4.52+1.62X}}$	*0.05*	*0.22*	*0.58*	*0.88*	*0.97*

Likelihood of data under our alternative hypothesis model =

$$.05^{1}.95^{49}.22^{13}.78^{37}.58^{29}.42^{21}.88^{44}.12^{6}.97^{48}.03^{2}$$

The log of this number (the Log Likelihood) is -95.38 :

THE VALUE REPORTED BY MINITAB

Definition : the DEVIANCE

The Deviance (denoted by G) is defined as

-2 * [log likelihood of null model – log likelihood of alternative model]

For tick data :

Deviance = G = -2 [-172.49 – (-95.38)] = 154.2*

Factoid : Under the null hypothesis, the Deviance G has a chi-square distribution with q degrees of freedom, where q is the number of parameters difference between the null model and alternative model.

For our data, q = 2-1 = 1 degree of freedom

SO : At Last we perform our hypothesis test!!!

P-value = prob (Observe 154.2 or larger value in a Chi-square distribution with 1 df) = 0.000

REJECT NULL HYPOTHESIS and conclude model contains significant predictors (i.e. concentration).

Goodness-of-Fit Tests

```
Goodness-of-Fit Tests

Method           Chi-Square  DF      P
Pearson             1.89804   3  0.594
Deviance            2.14244   3  0.543
Hosmer-Lemeshow     1.89804   3  0.594
```

Think about the following Hypothesis Test :

H_0 : Our Model fits the data well

H_a : Our Model does not fit the data well (there is a better model)

- Goodness of Fit tests are calculated by comparing the chosen model to the **saturated model**
- The **saturated model** is one that fits a separate probability of dying for each concentration level individually (i.e. just uses the sample probability of dying at each concentration level :

Log (Conc)	*1*	*2*	*3*	*4*	*5*
Sample Probability	*0.02*	*0.26*	*0.58*	*0.88*	*0.96*

- Calculate likelihood of saturated model :

$$.02^{1}.98^{49}.26^{13}.74^{37}.58^{29}.42^{21}.88^{44}.12^{6}.96^{48}.04^{2}$$

The log of this number (the Log Likelihood) is -94.3

- Calculate -2*[log likelihood of our model – log likelihood of saturated model]
- Under null hypothesis, this should have a chi-square distribution with 3 degrees of freedom
- *Degrees of freedom is number of parameters in saturated model (think 4 indicator variables) minus number of parameters in chosen model (1).*
- *MINITAB gives this test statistic as the **DEVIANCE** Goodness of Fit test. Null hypothesis is not rejected – i.e. there is no evidence that the saturated model fits any better than our model – **that is, our model seems to fit the data reasonably well!!!***

MINITAB provides observed and expected frequencies (calculated using model coefficients)

```
Table of Observed and Expected Frequencies:
(See Hosmer-Lemeshow Test for the Pearson Chi-Square Statistic)

                          Group
Value        1     2     3     4     5  Total
Success
  Obs        1    13    29    44    48    135
  Exp      2.6  10.8  29.1  43.8  48.6
Failure
  Obs       49    37    21     6     2    115
  Exp     47.4  39.2  20.9   6.2   1.4
Total       50    50    50    50    50    250
```

Measures of Association : not discussed here – if you want to know more, take Categorical Data Analysis!

Logisitic regression with categorical factors

***Example** : a study by Radelet (1981) examined 326 subjects eligible to receive the death penalty over a two-year period in Florida. Cases were classified by whether or not they received the death penalty, the race of the defendant, and the race of the victim.*

Defendant Race	Victim Race	Death Penalty	No Death Penalty	Pr(Death Penalty)
white	white	19	132	0.13
white	black	0	9	0.00
black	white	11	52	0.17
black	black	6	97	0.06

Here, we define receiving the death penalty as a ‘success’ only so that we can study increased risk of receiving the death penalty.

We fit the following logit model with factors for victim’s race and defendant’s race:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{vrace} + \beta_{drace} + \varepsilon$$

Logistic Regression with Categorical Factors in MINITAB : Just need to list categorical variables twice – once in model box, once in the FACTORS box.

Logistic Regression with Categorical Factors in SPSS : Click on the the Categorical Button to enter categorical data.

```
Logistic Regression Table
                                                 Odds       95% CI
Predictor        Coef     StDev       Z      P   Ratio    Lower     Upper
Constant      -2.8421    0.4203   -6.76  0.000
Defenden
 white        -0.4402    0.4009   -1.10  0.272    0.64     0.29      1.41
Victim
 white         1.3242    0.5193    2.55  0.011    3.76     1.36     10.40
```

For each factor, comparison is made to the category that is NOT reported *(i.e. in this case race=black).*

While the coefficient for defendant race is negative (i.e. white defendants are somewhat less likely to get the death penalty), this relationship is not significant (p-value 0.272).

However, the coefficient for victim race is significantly positive – that is, people who kill white people are significantly more likely to get the death penalty

```
Log-Likelihood = -109.541
Test that all slopes are zero: G = 7.431, DF = 2, P-Value = 0.024

Goodness-of-Fit Tests

Method                 Chi-Square    DF       P
Pearson                     0.376     1   0.540
Deviance                    0.701     1   0.403
Hosmer-Lemeshow             0.018     1   0.894
```

This indicates that the model has significant factors and that the model is a reasonable fit to the data.

This conclusion that race of victim and not race of defendant is a significant predictor of probability of receiving the death penalty has been found in numerous studies across states.

Example** : **Attitudes toward the death penalty.
Data is from some 1300+ people surveyed as part of the 1993 General Social Survey. Response is favor / do not favor the death penalty. Possible predictors are

- *Age (years)*
- *Political views (7 point scale)*
- *Gender*
- *Race (White, Black, Other)*
- *College Degree (yes / no)*

Analysis in class shows . . . . .