STATISTICS 538, Lecture #2

Logistic Regression

October 27, 2010

Space shuttle data: on i-th launch, y_i out of $n_i = 6$ 'O-Rings' sustained some damage, $i = 1, \dots, n = 23$

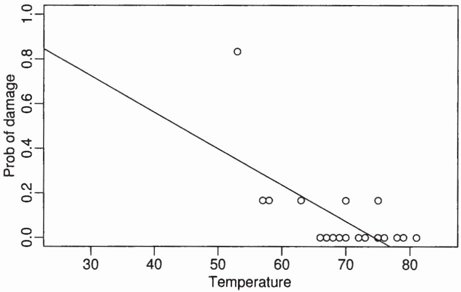

Figure 2.1 *Damage to O-rings in 23 space shuttle missions as a function of launch temperature. Least squares fit line is shown.*

Two equivalent data representations: R happy with either!

‘Bernoulli’		‘Binomial’		
OUTCOME	TEMP	‘S’	‘F’	TEMP

Logistic regression based on **logit** (actually log-odds!) function: $\text{logit}(p) = \log\{p/(1-p)\} = \log p - \log(1-p)$

$(Y_i|X^{(i)}) \sim \text{Binomial}(n_i, p_i)$,

where

$$\text{logit} p_i \quad = \quad \beta_0 + \beta_1 X_1^{(i)} + \ldots + \beta_{q-1} X_{q-1}^{(i)}$$

Note - completely specified - no variance to worry about.

Interpretation?

LM: β_j describes change in $E(Y|X)$ w.r.t. X_j, with X_{-j} held fixed.

Logistic regression?

Very natural if willing to think in terms of **odds**

e.g., Bernoulli data, X_j also binary:

$$\frac{\text{Odds}(Y = 1 | X_j = 1, X_{-j})}{\text{Odds}(Y = 1 | X_j = 0, X_{-j})} = \exp(\beta_j),$$

provided X_j not involved in interactions...

Continuous X_j?

This is the basis for interpreting logistic regression models, and explains the popularity of the logit function, rather than some other 'appropriate' function, for setting up a binary regression model.

Interpretation gets pushed a bit further, particularly in health sciences...

Mathematically, if p_a, p_b close to zero,

$$\frac{p_a/(1-p_a)}{p_b/(1-p_b)} \approx p_a/p_b$$

Mechanics of fitting: simple as $\hat{\beta} = (X^T X)^{-1} X^T Y$ for LM?

Not quite ...

Newton-Raphson to get MLE $\hat{\beta}$ (more later...)

Usual ML asymptotics to get $\hat{\text{Var}}\left(\hat{\beta}\right)$,
hence standard errors and confidence intervals.

R syntax: lm(y x, ...) $\rightarrow$ glm(y x, family=binomial, ...)

Shuttle data again

```
> logitmod <- glm(cbind(damage,6-damage) ~ temp, family=binomial, orings)
> summary(logitmod)
Deviance Residuals:
   Min       1Q  Median       3Q      Max
-0.953  -0.735  -0.439  -0.208   1.957

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.6630     3.2963    3.54    4e-04
temp         -0.2162     0.0532   -4.07   4.8e-05

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 16.912  on 21  degrees of freedom
AIC: 33.67
```

Goodness-of-fit (GOF) - easier than for LM!

Recall data structure $Y_i \sim \text{Binomial}(n_i, p_i)$, $i = 1, \ldots, n$.

Have (somehow) chosen regressors $(1, X_1, \ldots, X_{q-1})$.

Do likelihood ratio test:

H_0:

H_1:

Recall in general for a pair of **nested** models with likelihood functions $L_S(\theta_S)$ and $L_L(\theta_L)$:

Look at $D = 2\left\{\log L_L(\hat{\theta}_L) - \log L_S(\hat{\theta}_S)\right\}$.

If the smaller model is correct then

$D \overset{\text{approx}}{\sim} ?$

More on GOF

Shuttle data:

- no evidence of lack of fit for logit$p = \beta_0 + \beta_1$TEMP
- evidence of lack of fit for logit$p = \beta_0$

Very important caveat: conditions for the χ^2 approximation to the sampling distribution of the the deviance to be good?

Conceptual question: why can't the same approach to GOF work in the LM setting?