17/04/15.

MLA
LOGISTIC REGRESSION

- Parametric classification technique making distributional assumptions are able to allow for probabilistic quantification of class assignment

EXAMPLE
- Resting pulse depending on patient smoking and their weight
- Data 92 obs, 1=low, 0=high, Smokes 1=yes 0=no and weight in lbs

- P(Low) probability of low pulse rate
- To ensure P(low) falls within [0,1] we consider following model:

$$\text{Logit}(P(\text{Low})) = \text{Log}\left(\frac{P(\text{Low})}{1-P(\text{Low})}\right) = \alpha + \beta_s \text{Smokes} + \beta_w \text{Weight}$$

or $$\frac{P(\text{Low})}{1-P(\text{Low})} = \exp(\alpha + \beta_s \text{Smokes} + \beta_w \text{Weight})$$

or $$P(\text{Low}) = \frac{\exp(\alpha + \beta_s \text{Smokes} + \beta_w \text{Weight})}{1 + \exp(\alpha + \beta_s \text{Smokes} + \beta_w \text{Weight})}$$

VARIANTS
- Logistic model is not the only choice that guarantees $P(\text{Low}) \in [0,1]$
- Can use probit model:

$$\Phi^{-1}(P(\text{Low})) = \alpha + \beta_s \text{Smokes} + \beta_w \text{Weight}$$

Here $\Phi^{-1}(z) = w$ if probability of a normal (0,1) random variable is w then equal to z, it is the inverse CDF of $N(0,1)$

- Gompit / Complementary log-log: $-\text{Log}[-\log[P(\text{Low})]] = \alpha + \beta_s \text{Smokes} + \beta_w \text{Weight}$
- Known as LINK FUNCTIONS

- More evidence required in logistic model than the probit model to increase probability of assignment to that particular class

PROBABILITY
- N independent observations of blood pressure eg {L, H, L, L … H}
- What is probability of observing this?
- We know $P(\text{Low}) = \dfrac{\exp(\alpha + \beta_s \text{Smokes} + \beta_w \text{Weight})}{1 + \exp(\alpha + \beta_s \text{Smokes} + \beta_w \text{Weight})}$

- So $P(\text{High}) = \dfrac{1}{1 + \exp(\alpha + \beta_s \text{Smokes} + \beta_w \text{Weight})}$

- Also due to independence $P(L, H, H \ldots L)$ can be simplified to $P(L)\,P(L)\,P(H) \ldots P(H)$

- Can be expressed as

$$\prod_{i=1}^{n}\left[\frac{\exp(\alpha + \beta_s Smoked_i + \beta_w weight_i)}{1+\exp(\alpha + \beta_s Smoked_i + \beta_w weight_i)}\right]^{y_i} \times \prod_{i=1}^{n}\left[\frac{1}{1+\exp(\alpha + \beta_s Smoked_i + \beta_w weight_i)}\right]^{1-y_i}$$

- Here $y_i = 1$ if person is L and $y_i = 0$ if person = H.
- Find value of α, β_s, β_w that maximise the probabilities - MLE

INTREPRETING THE R OUTPUT
- Estimate column gives estimate of α, β_s, β_w. Sign important.
- 1 category = low, 0 = high
- Coef of Smoked is negative, if individual smoked they would have smaller probability of being assigned to the 1 category ⇒ more likely to have H rate
- Coef of weight is positive, heavier you weigh better chance of low.
- Be careful of categorical/binary covariates

Person who smoked and weighs 140lbs

$$\Rightarrow P(low) = \frac{\exp(-1.99 - 1.14\times1 + 0.03\times140)}{1+\exp(-1.99 - 1.19\times1 + 0.03\times140)} = 0.93$$

If person did not smoke ⇒ 0.98

STANDARD ERRORS
- SE record the uncertainty in the resulting estimate
- Tend to decrease with sample size
- 95% CI coef estimate ± 2 (SE of coef)
- $\beta_s \in (-2.30, -0.09)$, $\beta_w \in (0.001, 0.050)$
- Can be used to find 95% CI for $\exp(\beta_s)$ and $\exp(\beta_w)$
  $\exp(\beta_s) \in (0.1, 0.9)$ $\exp(\beta_w) \in (1.0, 1.05)$

TESTS
- May wish to test $H_0: \beta_s = 0$ v $H_1: \beta_s \neq 0$ or $H_0: \beta_w = 0$ v $H_1: \beta_w \neq 0$
- Compute $\frac{(\text{coef est}) - 0}{(\text{SE coef})}$
- Under H_0 this quantity is approximately Normal $(0,1)$

- If data had been standardised, then the magnitude of the coef would indicate the importance of the variable in predicting the class
- If we assume true coefficient β of a variable was 0, and that actual estimate generated $\hat{\beta}$ was random and subject to a draw from a normal dist with mean 0 and sd to estimated standard error $S\hat{\beta}$ then $\hat{\beta}/S\hat{\beta} \sim N(0,1)$

MLA

7/04/15 LOGISTIC REGRESSION

- The column z gives as $\hat{\beta}$ /SE and column $Pr(>|z|)$ tells us the probability of observing a z-value at least that far from 0 under the $N(0,1)$ distributional assumption

- If $Pr(>|z|)$ is small, then under the hyp that true coef is 0 and that the estimate is drawn from a normal distribution with mean 0 and sd as given by S.E., we have observed an unlikely outcome

- If $Pr(>|z|)$ is very small we may start to question the appropriateness of our original hypothesis, as we are faced with the dilemma of justifying a rare event.

- Note that $Pr(>|z|)$ is NOT the probability that the true coef is 0, it is the probability of having observed such a coef estimate under the assumption that the true coef is 0

- For resting pulse data, at 5% significance:
- The estimate for intercept α is not so extreme as to reject the hyp its true value is 0 - May consider model where its not included
- Estimate for β_S is extreme under H that its true value is 0, may wish to reject such H. If so, we may consider the Smokes Yes indicator as significant in predicting resting pulse
- Estimate for β_W is extreme under the H that its true value is 0, may reject such H. If so, may consider weight value as significant in predicting resting pulse

INTERACTIONS
- Including interactions adds possible products of the covariates to the model.
- In pulse data: $Log\left(\frac{P(Low)}{1-P(Low)}\right) = \alpha + \beta_S Smokes + \beta_W Weight + \beta_{SW} Smokes \times Weight$
- Interactions allow for effect of one covariate to be altered depending on value of another covariate
- In the above, the interaction would be reasonable if the effect that weight has upon resting pulse would differ depending on whether or not the individual smoked
- For example W of individual may be very important in predicting low pulse if individual smoked but of little importance if individual does not smoke
- When appropriate, interactions can greatly increase model's predictive ability

- Pick model with lowest AIC Akaike Information Criterion
$-2\log(\text{likelihood}) + 2p$

DEVIANCE

- Consider saturated model, same number of parameters as data points
- All 1 or 0.
- Likelihood is $L(\text{Data} \mid 1, \ldots, 1_n) = 1$
- Log likelihood is $l(\text{Data} \mid 1, \ldots, 1_n) = 0$

- Can be shown that under the hypothesis the proposed model is true, the deviance is approximately distributed as a χ^2_{m-k-1} distribution, where k is the number of parameters of proposed model

- Hence if deviance is > 95% point of the chi-square dist, we may claim under H the proposed model is true, we have obtained a rare event

NULL DEVIANCE

- Simplest model is one in which any data point is assigned a common probability of class assignment regardless of its covariate values
- In this case we have one parameter to determine probability of class assignment and it is estimated as the proportion of data in class "i"
- Deviance for proposed model is called null deviance
- If none of the covariates actually influence class assignment, the model deviance will not be much smaller than null deviance

$L_{null} \leq L_{mod} \leq L_{max}$.

- Deviance (residual dev) = $2[\log(L_{max}) - \log(L_{mod})]$
- null dev = $2[\log(L_{max}) - \log(L_{null})]$

- For n reasonably large and m small, can interpret residual dev as a measure of fit.
- Calculate 1 - pchisq (resid.deviance, deg.freedom)
- If resulting value is v small, then a rare event has occured under the assumption that our proposed model is correct.

- To test if model is explaining variation in the data we can compute the p-value from a goodness of fit test.

1 - pchisq (deviance(res), df.residual(res))

- Smaller the value, greater the evidence that there is a significant difference between full model and model of interest. 0.05 cut off assumed
- Often will be a significant difference, so checking the difference of deviance against null deviance against a χ^2_p-dist will indicate whether the model of interest is better than the null model with no covariates

MLA

.../15 LOGISTIC REGRESSION

- Considered as a classification technique when there are two groups
- A new observation is classified as being of one group or another depending on whether the predicted probability falls above or below the threshold of 0.5.
- Similar to LDA.

- In LDA, decision boundary between class k and class l given by

$$\log \frac{P(k|x)}{P(l|x)} = \log \frac{\pi_k}{\pi_l} + \log \frac{f(x|k)}{f(x|l)} = 0$$

- In LR, by assumption: $\log \frac{P(k|x)}{P(l|x)} = \beta_0 + \beta^T x$

- Hence models have same form
- Difference lies in way linear coefficients are estimated