Applied Regression

Generalized Linear Model - GLM - Part-2

Module 9 Lecture - 9-2

Testing Hypotheses on Subsets of Parameters:

We can also use the deviance to test hypotheses on subsets of the model parameters, just as we used the difference in regression (or error) sums of squares to test similar hypotheses in the normal-error linear regression model case. Recall that the model can be written as

$$\eta = X\beta = X_1\,\beta_1 \,+\, X_2\,\beta_2$$

where the full model has p parameters, β_1 contains (p - r) of these parameters, β_2 contains r of these parameters, and the columns of the matrices X_1 and X_2 contain the variables associated with these parameters.

The deviance of the full model will be denoted by $D(\beta)$. Suppose that we wish to test the hypotheses

$$H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$$

Therefore, the reduced model is

$$\eta = X_1\,\beta_1$$

Now fit the reduced model, and let $D(\beta_1)$ be the deviance for the reduced model.

The deviance for the reduced model will always be larger than the deviance for the full model, because the reduced model contains fewer parameters. However, if the deviance for the reduced model is not much larger than the deviance for the full model, it indicates that the reduced model is about as good a fit as the full model, so it is likely that the parameters in β_2 are equal to zero. That is, we cannot reject the null hypothesis above. However, if the difference in deviance is large, at least one of the parameters in β_2 is likely not zero, and we should reject the null hypothesis. Formally, the difference in deviance is

$$D(\beta_2|\beta_1) = D(\beta_1) - D(\beta)$$

and this quantity has n - (p - r) - (n - p) = r degrees of freedom. If the null hypothesis is true and if n is large, the difference in deviance has a chi-square distribution with r degrees of freedom.

Therefore, the test statistic and decision criteria are

$$\begin{aligned} &\text{if} \quad D(\beta_2|\beta_1) \geq \chi^2_{\alpha,r} \quad \text{reject} \;\; H_0 \\ &\text{if} \quad D(\beta_2|\beta_1) < \chi^2_{\alpha,r} \quad \text{do not reject} \;\; H_0 \end{aligned}$$

Sometimes the difference in deviance $D(\beta_2|\beta_1)$ is called the partial deviance.

Revisiting the Previous Problem: Suppose that we wish to determine whether adding a quadratic term in the linear predictor would improve the model. Therefore, we will consider the full model to be

$$y = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x + \beta_2 x^2)}}$$

Now the linear predictor for the full model can be written as

$$\eta = X\,\beta = \beta_0 + \beta_1\,x + \beta_2\,x^2 = X_1\,\beta_1 + X_2\,\beta_2$$

Reduced Model: $\eta = \beta_0 + \beta_1\,x$ and Full model: $\eta = \beta_0 + \beta_1\,x + \beta_2\,x^2$

Hence we want to test $H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$.

We find that the deviance for the full model is

$D(\beta) = 3.28164$ with n - p = 8 - 3 = 5 degrees of freedom.

Now the reduced model (i.e our original model) $D(\beta_1) = 6.05077$

Hence the difference

$D(\beta_2|\beta_1) = D(\beta_1) - D(\beta) = 6.0577 - 3.28164 = 2.76913$

This statistic is supposed to follow Chi-Square distribution with r = 1 degrees of freedom.

Since the P value associated with the difference in deviance is 0.0961, hence at 5% level we cannot reject H_0 but we might conclude that there is some marginal value in including the quadratic term in the regressor variable x = years of exposure in the linear predictor for the logistic regression model.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr ChiSq
Deviance	3.2816	5	0.6563	0.6567
Pearson	2.9445	5	0.5889	0.7085
HL	2.8024	5		0.7304

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	SE	Chi-Sqr	p-value
Intercept	1	-6.7099	1.5349	19.1096	<.0001
Years of Exposure	1	0.2276	0.0927	6.0205	0.0141
Year SQ	1	-0.00208	0.00136	2.3314	0.1268

Remark: Notice that the P value for β_2 is P = 0.127, suggesting that the squared term in years of exposure does not contribute significantly to the fit. Recall from the deviance calculation previously, we tested for the significance of β_2 using the partial deviance method we obtained a different P value (0.0961). Now in MLR model, the t test on a single regressor is equivalent to the partial F test on a single variable (recall that the square of the t statistic is equal to the partial F statistic). However, this equivalence is only true for linear models, and the GLM is a nonlinear model. Here we are using two different test statistic to do the same test.

Confidence Interval for Odds Ratio

Reconsider the original logistic regression model that we fit to the pneumoconiosis data. The estimated model is

$$\hat{y} = \hat{\pi} = \frac{1}{1 + e^{4.7965 - 0.0935\, x}}$$

We found (in last lecture) the 95% CI for β_1 is (0.0632, 0.1237) .

As Odds Ration(OR) $= e^{\beta_1}$, now we can find a 95% CI for OR as follows:

$$e^{0.0632} < OR < e^{0.1237} \quad \Rightarrow \quad 1.07 < OR < 1.13$$

Remark: Even though the estimated OR=1.10 is almost the mid point of the interval, it is not necessary to happen all the time.

Confidence Intervals for linear predictor

It is possible to find a CI for the linear predictor at any set of values of the predictor variables that is of interest.

Let $x_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$ be the values of the regressor variables that are of interest. The linear predictor evaluated at x_0 is $x_0' \hat{\beta}$.

The variance of the linear predictor at this point is

$$Var(x_0'\,\hat{\beta}) = x_0'\,Var(\hat{\beta})\,x_0 = x_0'\,(X'VX)^{-1}\,x_0$$

So the $100(1-\alpha)$ percent CI on the linear predictor is

$$x_0'\,\hat{\beta} \;-\; z_{\alpha/2}\,\sqrt{x_0'\,(X'VX)^{-1}\,x_0} \leq x_0'\,\beta \;\leq x_0'\,\hat{\beta} \;+\; z_{\alpha/2}\,\sqrt{x_0'\,(X'VX)^{-1}\,x_0}$$

Consequently, the CI on the linear predictor given above enables us to find a CI on the estimated probability of success π_0 at the point of interest $x_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}$. Let

$$L(x_0) = x_0'\,\hat{\beta} \;-\; z_{\alpha/2}\,\sqrt{x_0'\,(X'VX)^{-1}\,x_0}$$

and

$$U(x_0) = x_0'\,\hat{\beta} \;+\; z_{\alpha/2}\,\sqrt{x_0'\,(X'VX)^{-1}\,x_0}$$

are the lower and upper confidence limit for the linear predictor.

Now the point prediction of the probability of success is $\hat{\pi} = \frac{exp(x_0'\,\hat{\beta})}{1+exp(x_0'\,\hat{\beta})}$ Hence the $100(1-\alpha)\%$ confidence interval is

$$\frac{exp[L(x_0)]}{1 + exp[L(x_0)]} \leq \pi \leq \frac{exp[U(x_0)]}{1 + exp[U(x_0)]}$$

Revisiting the Example:

Suppose that we want to find a 95% CI on the probability of miners with x = 40 years of exposure contracting pneumoconiosis. From the fitted logistic regression model, we can calculate a point estimate of the probability at 40 years of exposure as

$$\hat{\pi}_0 = \frac{1}{1 + e^{4.7965 - 0.0935\,(40)}} = 0.2580$$

To find the CI, we need to calculate the variance of the linear predictor at this point. The variance is

$$\begin{aligned} Var(x_0'\,\hat{\beta}) &= x_0'\,(X'VX)^{-1}\,x_0 \\ &= (1\,40)\begin{pmatrix} 0.32330 & -0.008348 \\ -0.008348 & 0.000238 \end{pmatrix}\begin{pmatrix} 1 \\ 40 \end{pmatrix} = 0.036243 \end{aligned}$$

Now

$$L(x_0) = x_0' \hat{\beta} - z_{\alpha/2} \sqrt{x_0' (X'VX)^{-1} x_0}$$
$$= -4.7965 + 0.0935 - 1.96\sqrt{0.036243} = -1.4296$$

and

$$U(x_0) = x_0' \hat{\beta} + z_{\alpha/2} \sqrt{x_0' (X'VX)^{-1} x_0}$$
$$= -4.7965 + 0.0935 + 1.96\sqrt{0.036243} = -0.06834$$

Therefore the 95% CI on the estimated probability of contracting pneumoconiosis for miners that have 40 years of exposure is

$$\frac{exp[L(x_0)]}{1 + exp[L(x_0)]} \leq \pi_0 \leq \frac{exp[U(x_0)]}{1 + exp[U(x_0)]}$$

$$\frac{exp[-1.4296]}{1 + exp[-1.4296]} \leq \pi_0 \leq \frac{exp[-0.06834]}{1 + exp[-0.06834]}$$

$$0.1932 \leq \pi_0 \leq 0.3355$$

Problem: The compressive strength of an alloy fastener used in aircraft construction is being studied. Ten loads were selected over the range 2500 - 4300 psi and a number of fasteners were tested at those loads. The numbers of fasteners failing at each load were recorded. The complete test data are shown below.

a. Fit a logistic regression model to the data. Use a simple linear regression model as the structure for the linear predictor.

Using the SAS output we see the predicted model is

$$\hat{y} = \frac{1}{1 + e^{-(-5.3397 + 0.00155\,x)}}$$

b. Does the model deviance indicate that the logistic regression model from part a is adequate?

The Deviance = 0.3719 with df = 8. It also provides the p-value =1.00. Hence the model is fit is good as p-value is high and also D/df $<<$ 1.00.

Remark: Pearson and HL stat also shows a good fit.

c. Expand the linear predictor to include a quadratic term. Is there any evidence that this quadratic term is required in the model?

Second model was tried with the addition of the quadratic term and deviance is = 0.2837.

Hence to test $H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$ using deviance we observe

Difference in deviance = 0.3819 - 0.2837 = 0.0982 $< \chi^2_{1,0.05}$. Hence it is not significant.

d. For the quadratic model in part c, find Wald statistics for each individual model parameter.

From SAS output we that for testing $H_0 : \beta_1 = 0 \quad vs \quad H_1 : \beta_1 \neq 0$, we see that Chi-Square statistic = 0.1746 and p-value = 0.6761. Hence it is not significant.

Also from SAS output we that for testing $H_0 : \beta_2 = 0 \quad vs \quad H_1 : \beta_2 \neq 0$, we see that Chi-Square statistic = 0.0882 and p-value = 0.7665. Hence it is also not significant.

Remark: But the first model had β_1 significant.