STAT 350: Lecture 35

Estimating equations: an introduction via glm

Estimating Equations: refers to equations of the form

$$h(X, \theta) = 0$$

which are solved for θ to get estimates $\hat{\theta}$. Examples:

1. The normal equations in linear regression:

$$X^T Y - X^T X\beta = 0$$

2. The likelihood equations:

$$\frac{\partial \ell}{\partial \theta} = 0$$

   where $\ell(\theta)$ is the log-likelihood.

3. The equation which must be solved to do non-linear least squares:

$$\sum (Y_i - \mu_i)\frac{\partial \mu_i}{\partial \theta} = 0$$

4. The iteratively reweighted least squares estimating equation:

$$\sum \frac{Y_i - \mu_i}{\sigma_i^2}\frac{\partial \mu_i}{\partial \theta} = 0$$

   where, in a generalized linear model the variance σ_i^2 is a *known* (except possibly for a multiplicative constant) function of μ_i.

Only the first of these equations can usually be solved analytically. In Lecture 34 I showed you an example of an iterative technique of solving such equations.

Theory of Generalized Linear Models

The likelihood function for a Poisson regression model is:

$$L(\beta) = \prod \frac{\mu_i^{y_i}}{y_i!} \exp(-\sum \mu_i)$$

and the log-likelihood is

$$\sum y_i \log \mu_i - \sum \mu_i - \sum \log(y_i!)$$

A typical glm model is

$$\mu_i = \exp(x_i^T \beta)$$

where the x_i are covariate values for the ith observation (often including an intercept term just as in standard linear regression).

In this case the log-likelihood is

$$\sum y_i x_i^T \beta - \sum \exp(x_i^T \beta) - \sum \log(y_i!)$$

which should be treated as a function of β and maximized.

The derivative of this log-likelihood with respect to β_k is

$$\sum y_i x_{ik} - \sum \exp(x_i^T \beta) x_{i,k} = \sum (y_i - \mu_i) x_{i,k}$$

If β has p components then setting these p derivatives equal to 0 gives the **likelihood equations**.

For a Poisson model the variance is given by

$$\sigma_i^2 = \mu_i = \exp(x_i^T \beta)$$

so the likelihood equations can be written as

$$\sum \frac{(y_i - \mu_i) x_{i,k} \mu_i}{\mu_i} = \sum \frac{(y_i - \mu_i)}{\sigma_i^2} \frac{\partial \mu_i}{\partial \beta_k} = 0$$

which is the fourth equation above.

These equations are solved iteratively, as in non-linear regression, but with the iteration now involving *weighted* least squares. The resulting scheme is called **iteratively reweighted least squares**.

1. Begin with a guess for the standard deviations σ_i (taking them all equal to 1 is simple).

2. Do (non-linear) weighted least squares using the guessed weights. Get estimated regression parameters $\hat\beta^{(0)}$.

3. Use these to compute estimated variances $\hat\sigma_i^2$. Go back to do weighted least squares with these weights and get $\hat\beta^{(1)}$.

4. Iterate (repeat over and over) until estimates not really changing.

If the $\hat\beta^{(k)}$ converge as $k \to \infty$ to something, say, $\hat\beta$ then since

$$\sum \left[\frac{y_i - \mu_i(\hat\beta^{(k+1)})}{\sigma_i^2(\hat\beta^{(k)})} \right] \frac{\partial \mu_i(\hat\beta^{(k+1)})}{\partial \hat\beta^{(k+1)}} = 0$$

we learn that $\hat\beta$ must be a root of the equation

$$\sum \left[\frac{y_i - \mu_i(\hat\beta)}{\sigma_i^2(\hat\beta)} \right] \frac{\partial \mu_i(\hat\beta)}{\partial \hat\beta} = 0$$

which is the last of our example estimating equations.

Distribution of Estimators

Distribution Theory is the subject of computing the distribution of statistics, estimators and pivots. Examples in this course are the Multivariate Normal Distribution, the theorems about the chi-squared distribution of quadratic forms, the theorems that F statistics have F distributions when the null hypothesis is true, the theorems that show a t pivot has a t distribution.
Exact Distribution Theory: name applied to exact results such as those in previous example when the errors are assumed to have *exactly* normal distributions.
Asymptotic or **Large Sample Distribution Theory**: same sort of conclusions but only approximately true and assuming n is large. Theorems of the form:

$$\lim_{n\to\infty} P(T_n \le t) = F(t)$$

- An estimate is normally only useful if it is equipped with a measure of uncertainty such as a standard error.

- A standard error is a useful measure of uncertainty provided the error of estimation $\hat\theta - \theta$ has approximately a normal distribution and the standard error is the standard deviation of this normal distribution.

- For many estimating equations $h(Y, \theta) = 0$ the root $\hat\theta$ is unique and has the desired approximate normal distribution, **provided the sample size n is large**.

Sketch of reasoning in special case

POISSON EXAMPLE: $p = 1$

Assume Y_i has a Poisson distribution with mean $\mu_i = e^{x_i\beta}$ where now β is a scalar.

The estimating equation (the likelihood equation) is

$$U(\beta) = h(Y_1, \ldots, Y_n, \beta) = \sum (Y_i - e^{x_i\beta})x_i = 0$$

It is now important to distinguish between a value of β which we are trying out in the estimating equation and the true value of β which I will call β_0. If we happen to try out the true value of β in U then we find

$$E_{\beta_0}(U(\beta_0)) = \sum x_i E_{\beta_0}(Y_i - \mu_i) = 0$$

On the other hand if we try out a value of β other than the correct one we find

$$E_{\beta_0}(U(\beta)) = \sum x_i(e^{x_i\beta} - e^{x_i\beta_0}) \neq 0\,.$$

But $U(\beta)$ is a sum of independent random variables so by the law of large numbers (law of averages) must be close to its expected value. This means: if we stick in a value of β far from the right value we will not get 0 while if we stick in a value of β close to the right answer we will get something close to 0. This can sometimes be turned in to the assertion:
The glm estimate of β is **consistent**, that is, it converges to the correct answer as the sample size goes to ∞.

The next theoretical step is another **linearization**. If $\hat\beta$ is the root of the equation, that is, $U(\hat\beta) = 0$, then

$$0 = U(\hat\beta) \approx U(\beta_0) + (\hat\beta - \beta_0)U'(\beta_0)$$

This is a **Taylor's expansion**. In our case the derivative U' is

$$U'(\beta) = -\sum x_i^2 e^{x_i\beta}$$

so that approximately

$$\hat\beta = \frac{\sum(Y_i - \mu_i)x_i}{\sum x_i^2 e^{x_i\beta_0}}$$

The right hand side of this formula has expected value 0, variance

$$\frac{\sum x_i^2 Var(Y_i)}{\left(\sum x_i^2 e^{x_i\beta_0}\right)^2}$$

which simplifies to

$$\frac{1}{\sum x_i^2 e^{x_i\beta_0}}$$

This means that an approximate standard error of $\hat\beta$ is

$$\frac{1}{\sqrt{\sum x_i^2 e^{x_i\beta_0}}}$$

that an estimated approximate standard error is

$$\frac{1}{\sqrt{\sum x_i^2 e^{x_i\hat\beta}}}$$

Finally, since the formula shows that $\hat\beta - \beta_0$ is a sum of independent terms the central limit theorem suggests that $\hat\beta$ has an approximate normal distribution and that

$$\sqrt{\sum x_i^2 e^{x_i\hat\beta}}(\hat\beta - \beta_0)$$

is an approximate pivot with approximately a $N(0,1)$ distribution. You should be able to turn this assertion into a 95% (approximate) confidence interval for β_0.

Scope of these ideas

The ideas in the above calculation can be used in many contexts.

- We can get approximate standard errors in non-linear regression.
- We can get approximate standard errors in any model where we do maximum likelihood.
- We can show that the assumption of normal errors does not have too big an impact on the t and F tests in multiple regression.
- We can get approximate standard errors in generalized linear models.

- We can demonstrate that the role of the Error Sum of Squares in multiple regression can be replaced, approximately, by a function called the **Deviance** which is a function whose *derivative* (with respect to the parameters) is the estimating equation.

Further exploration of the ideas in this course

- **STAT 402** explores applications of generalized models.
- **STAT 410** applies regression to samples from finite populations.
- **STAT 420** discusses the analysis of variance and regression when the normality assumption seems very probably wrong.
- **STAT 430** discusses the design and analysis of experiments. Topics include: designs which save on effort by deliberately making $X^T X$ singular, the advantages of randomized controlled experiments, justification of t and F tests by randomization rather than sampling arguments, experiments in which (some of) the βs are random.
- **STAT 450** discusses the exact and approximate distribution theory discussed here, along with mathematical justifications for using tests and estimates recommended here and in other courses as opposed to any others.
- **STAT 802** explores problems in which each Y_i is **multivariate**, that is, there is more than 1 response variable.
- **STAT 804** (Time Series Analysis) explores problems in which the errors ϵ_i are not independent.