Maximum Likelihood Estimation

Charles H. Franklin

franklin@polisci.wisc.edu

University of Wisconsin - Madison

Lecture 5

Normal Regression

Last Modified: June 13, 2005

Normal Regression

- Let's begin with the idea that our dependent variable is normally distributed. That implies it is continuous and unbounded.
- $$f(y_i|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}}$$
- As written, this distribution says that each observation, y_i is drawn from a normal distribution with constant mean, μ and constant variance, σ^2.
- This is only useful if all we care to model is the marginal distribution of y_i and are willing to assume constant mean and variance.

Normal Regression

- But for a regression model, we explicitly reject this assumption. We assume that the mean of y_i varies, and we wish to model this variation in the mean.
- Assume y_i is distributed normally:

$$f(y_i|\mu_i,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{1}{2}\frac{(y_i-\mu_i)^2}{\sigma^2}}$$

- The only notational change is to add a subscript i to μ. But substantively, this changes everything. Now each observation is drawn from a *different* normal distribution with a common variance but with a unique mean, μ_i.
- It may not be obvious, but the elemental substance has just impacted the rotating blade.

Identification

- What is the problem with this model?
- There are now $N+1$ parameters to be estimated from N observations!
- So if we draw $y_1 = 200$ and $y_2 = -50$, we cannot know if this is because both draws are from a distribution with mean 75 and standard deviation 125, which would make both observations pretty likely, OR if y_1 comes from a distribution with mean 199 and standard deviation 1, while y_2 is drawn from a normal with mean -49 and standard deviation 1.

Identification

- Things are even worse for estimating the variance. Since μ_i varies from observation to observation, the sample variance confounds variation in μ_i with the common variance of the distributions, σ^2. There is no way to disentangle these sources of variation.
- While μ_i may be identified (trivially), as $\mu_i = y_i$, it is not possible to identify the variance parameter, σ^2.
- Yet the alternative to this specification, claiming that all observations have a common mean, is equally unappealing for it homogenizes everything, allowing for no meaningful differences, only chance variation.

Reparameterization

- If it is substantively reasonable to do so, we can reparameterize μ_i in terms of a small number of new parameters and some observed exogenous variables.
- For example, $\mu_i = x_i\beta$, where x_i is a $1 \times k$ vector and β is $k \times 1$.
- In this way, we reduce the n parameters of μ_i into a mere k parameters in β.

Reparameterization

- The trick, of course, is in replacing unexplained variation in μ_i with observed exogenous variables and a small number of parameters.
- By observing variation in x_i, and by linking this variation to variation in μ_i, we are able to allow each observation to be drawn from a distribution with a unique mean.
- Unlike before, however, now we can explain this variation in terms of other, observable, characteristics.

Reparameterization

- The invariance property of ML estimators makes this possible. If we estimate $\hat{\beta}$ and from this $\hat{\mu}_i = x_i\hat{\beta}$ then $\hat{\mu}_i$ is also the ML estimator of μ_i by the invariance property.
- Does this reparameterization make substantive sense? Because the mean of a normal is unbounded and continuous, the linear function of x_i is appropriate.
- If μ_i were a priori known to be positive, then this would not be a very good choice of parameterization, since for a sufficiently negative value of x_i, the function $x_i\beta$ would also be negative, implying a value of μ_i outside its a priori range.
- If the normal distribution makes sense, then there is little reason to believe that μ_i should have a restricted range. Hence the reparameterization makes sense.

Normal Regression

- The distribution for each case now becomes:

$$y_i \sim \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}}$$

- The likelihood for the sample is

$$L(y|x,\beta,\sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}}$$

Normal Regression

- and the log-likelihood is

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y_i - x_i\beta)^2}{\sigma^2}}\right) \\ &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{N}\left[\frac{(y_i - x_i\beta)^2}{\sigma^2}\right] \end{aligned}$$

Finding the ML Estimator

- In this case, we can find a closed form solution for the parameters (β, σ^2).
- First, rewrite the sum of the log-likelihood in matrix form:

$$\begin{aligned} \ln L(y|x,\beta,\sigma^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) \\ &\quad -\frac{1}{2}\left[\frac{(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)}{\sigma^2}\right] \end{aligned}$$

  where $\mathbf{y}$ is an $N \times 1$ vector and $\mathbf{x}$ is a $N \times k$ matrix.
- (You should study this formula until you convince yourself that this actually is the expression for the sum of the log-likelihood.)

Finding the ML Estimator

- Expanding the numerator of the last term, and moving the scalar σ^2 out front gives

$$\begin{aligned} \ln L(y|x,\beta,\sigma^2) &= -\frac{N}{2}\ln(2\pi) - \frac{N}{2}\ln(\sigma^2) \\ &\quad -\frac{1}{2\sigma^2}\left[\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{x}'\mathbf{y} + \beta'\mathbf{x}'\mathbf{x}\beta\right] \end{aligned}$$

- Now take the derivative of this sum of the log-likelihood to get

$$\frac{\partial \ln L}{\partial \beta} = -\frac{1}{2\sigma^2}\left[\frac{\partial[\mathbf{y}'\mathbf{y} - 2\beta'\mathbf{x}'\mathbf{y} + \beta'\mathbf{x}'\mathbf{x}\beta]}{\partial \beta}\right]$$

Finding the ML Estimator

- Remember that β is a vector, and apply the rules for matrix differentiation to get

$$\frac{\partial \ln L}{\partial \beta} = -\frac{1}{2\sigma^2}\left[-2\mathbf{x}'\mathbf{y} + 2\mathbf{x}'\mathbf{x}\beta\right]$$

- After factoring the -2 and cancelling, we have

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2}\left[\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta\right]$$

Finding the ML Estimator

- Set this equal to zero, and the rest is easy:

$$\begin{aligned}\frac{1}{\sigma^2}\left[\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta\right] &= 0\\ \mathbf{x}'\mathbf{x}\beta &= \mathbf{x}'\mathbf{y}\\ \hat{\beta} &= (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}\end{aligned}$$

- And this is, of course, the familiar formula for an OLS coefficient vector. As I promised, OLS and ML give the same estimator for the coefficients.

Finding the ML Estimator

- What about the variance, σ^2? Take the derivative of the log-likelihood wrt σ^2:

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)\right]$$

- Setting this to zero and carrying out the algebra gives:

$$\begin{aligned}-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\left[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)\right] &= 0\\ \frac{1}{2\sigma^4}\left[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)\right] &= \frac{N}{2\sigma^2}\\ \frac{1}{\sigma^2}\left[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)\right] &= N\end{aligned}$$

Finding the ML Estimator

- Since we've solved for $\hat{\beta}$ we can solve this for σ^2 by replacing β with its estimate:

$$\begin{aligned}\frac{1}{\sigma^2}\left[(\mathbf{y} - \mathbf{x}\hat{\beta})'(\mathbf{y} - \mathbf{x}\hat{\beta})\right] &= N\\ \frac{1}{\sigma^2}\left[(\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}})\right] &= N\\ \frac{1}{\sigma^2}\left[\mathbf{e}'\mathbf{e}\right] &= N\\ \hat{\sigma}^2 &= \frac{\mathbf{e}'\mathbf{e}}{N}\end{aligned}$$

- Which is NOT the OLS solution, but which is clearly asymptotically equivalent ($\hat{\sigma}^2_{ols} = \frac{\mathbf{e}'\mathbf{e}}{N-k}$).

Variance-Covariance Matrix

- We can also find the Hessian for this model by taking second derivatives. Let $\theta' = [\beta' \ \sigma^2]$. Then

$$\frac{\partial \text{lnL}}{\partial \theta} = \begin{bmatrix} \frac{1}{\sigma^2}[\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta] \\ -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)] \end{bmatrix}$$

  is a $(k+1) \times 1$ vector.
- We now want the derivative of this vector, wrt θ', which will produce a $(k+1) \times (k+1)$ matrix, namely the Hessian.

Variance-Covariance Matrix

- $$\frac{\partial^2 \text{lnL}}{\partial\theta\partial\theta'} = \begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{x}'\mathbf{x} & -\frac{1}{\sigma^4}[\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta] \\ -\frac{1}{\sigma^4}[\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta] & \frac{N}{2\sigma^4} - \frac{1}{\sigma^6}[(\mathbf{y} - \mathbf{x}\beta)'(\mathbf{y} - \mathbf{x}\beta)] \end{bmatrix}$$
- Check the dimensions of each submatrix and be sure you understand why they are as they are.
- The last row and last column are less familiar, representing the covariance of σ^2 with each element in β and the variance of the estimate of σ^2.

The Information Matrix

- The information matrix is the negative of the expected value of the Hessian. Taking $\mathbf{x}$ as fixed and using the fact that $\text{E}(\mathbf{y}) = \mathbf{x}\beta$ so $\text{E}(\mathbf{x}'\mathbf{y} - \mathbf{x}'\mathbf{x}\beta) = 0$, we can find the expected value:
-

$$\begin{aligned} I(\theta) &= -E[H(\theta)] \\ &= -\begin{bmatrix} -\frac{1}{\sigma^2}\mathbf{x}'\mathbf{x} & 0 \\ 0 & -\frac{N}{2\sigma^4} \end{bmatrix} \end{aligned}$$

The Information Matrix

- The inverse of the information matrix is the variance covariance matrix of the ML parameter estimates, which in this case is simply
-

$$I(\theta)^{-1} = \begin{bmatrix} \sigma^2(\mathbf{x}'\mathbf{x})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{N} \end{bmatrix}$$

Recap

- We can write the normal regression model, which we usually estimate via OLS, as a ML model.
- Write down the log-likelihood
- Take derivatives wrt the parameters
- Set the derivatives to zero
- Solve for the parameters

Recap

- We get an estimator of the coefficient vector which is identical to that from OLS.
- The ML estimator of the variance, however, is different from the least squares estimator. The reason for the difference is that the OLS estimator of the variance is unbiased, while the ML estimator is biased but consistent. In large samples, as assumed by ML, the difference is insignificant.

Recap

- Finally, we can apply the formula for the information matrix to get the variance-covariance matrix of the ML parameters.
- This turns out to give the familiar formula for the variance-covariance matrix of the parameters, $\sigma^2(\mathbf{x}'\mathbf{x})^{-1}$
- and a simple, if unfamiliar expression for the variance of $\hat{\sigma}^2$.

Inference

- Since the parameter estimates are all MLEs, they are all asymptotically normally distributed.
- The square root of the diagonal elements of the inverse of the information matrix gives us estimates of the standard errors of the parameter estimates.
- We can construct a simple z-score to test the null hypothesis concerning any individual parameter, just as in OLS, but using the normal instead of the t-distribution.

Inference

- Though we have not yet developed it, we can also construct a likelihood ratio test for the null hypothesis that all elements of β except the first are zero .
- This corresponds to the F-test in a least squares model, a test that none of the independent variables have an effect on y.

Using the ML Normal Regression

- So should you now use this to estimate normal regression models?
- *No, of course not!*
- Because OLS is unbiased regardless of sample size.
- There is an enormous amount of software available to do OLS.
- Since the ML and OLS estimates are asymptotically identical, there is no gain in switching to ML for this standard problem.

Not a waste of time!

- Now seen a fully worked, non-trivial, application of ML to a model you are already familiar with.
- But there is a much better reason to understand the ML model for normal regression:
- Once the usual assumptions no longer hold, ML is easily adapted to the new case, something that cannot be said of OLS.