Stat243: section 12 practice problem

November 22, 2015

1. Consider a censored regression problem. We assume a simple linear regression model, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$. Suppose we have an iid sample, but that for any observation with $Y > \tau$, all we are told is that Y exceeded the threshold and not its actual value. In a given sample, c of the n observations will (in a stochastic fashion) be censored, depending on how many exceed the fixed τ. A real world example (but with truncation in the left tail) is in measuring pollutants, for which values below a threshold are reported as below the limit of detection.

   (a) Design an EM algorithm to estimate the 3 parameters, $\theta = (\beta_0, \beta_1, \sigma^2)$, taking the complete data to be the available data plus the actual values of the truncated observations. You'll need to make use of $E(Y|Y > \tau)$ and $\text{Var}(Y|Y > \tau)$ where Y is normally distributed. Be careful that you carefully distinguish θ from the current value at iteration t, θ_t, in writing out the expected log-likelihood and computing the expectation and that your maximization be with respect to θ. You should be able to analytically maximize the expected log likelihood. A couple hints:

      i. From the Johnson and Kotz bibles on distributions, the mean and variance of the truncated normal distribution, $f(Y) \propto \mathcal{N}(\mu, \sigma^2) I(Y > \tau)$, are:

      $$\begin{aligned} E(Y|Y > \tau) &= \mu + \sigma\rho(\tau^*) \\ V(Y|Y > \tau) &= \sigma^2 \left(1 + \tau^*\rho(\tau^*) - \rho(\tau^*)^2\right) \\ \rho(\tau^*) &= \frac{\phi(\tau^*)}{1 - \Phi(\tau^*)} \\ \tau^* &= (\tau - \mu)/\sigma, \end{aligned}$$

      where $\phi(\cdot)$ is the standard normal density and $\Phi(\cdot)$ is the standard normal CDF.

      ii. You should recognize that your expected log-likelihood can be expressed as a regression of $\{Y_{obs}, m_t\}$ on $\{x\}$ where Y_{obs} are the non-censored data and $\{m_{i,t}\}$, $i = 1, \ldots, c$ are used in place of the censored observations. Note that $\{m_{i,t}\}$ will be functions of θ_t and thus constant in terms of the maximization step. Your estimator for σ^2 should involve a ratio where the numerator involves the usual sum of squares for the non-censored data plus two additional terms that you should interpret statistically.

   (b) Propose reasonable starting values for the 3 parameters as functions of the observations.

   (c) Write an R function, with auxiliary functions as needed, to estimate the parameters. Make use of the initialization from part (b). You may use *lm()* for updating β. You'll need to include criteria for deciding when to stop the optimization. Test your function using data simulated from the model with (a) a modest proportion of exceedances expected, say 20%, and (b) a high proportion, say 80%. Take $n = 100$ and the parameters such that with complete data, $\hat{\beta}_1/se(\hat{\beta}_1) \approx 3$. (In other words, you'll need to figure out values of β_1 and σ^2 such that the signal to noise ratio is 3.) You'll also need to generate the xs in some reasonable fashion.

(d) A different approach to this problem just directly maximizes the log-likelihood of the observed data, which for the censored observations just involves the likelihood terms, $P(Y_i > \tau)$. Estimate the parameters (and standard errors) for your test cases using *optim()* with the BFGS option in R. You will want to consider reparameterization, and possibly use of the *parscale* argument. Compare how many iterations EM and BFGS take. Note that parts (c) and (d) together provide a nice test of your code.