Problem Set 8

Due Monday Dec. 4, 5 pm

Comments

- This covers material in Unit 11.
- It's due at 10 am (Pacific) on December 4, both submitted as a PDF to Gradescope as well as committed to your GitHub repository.
- Please see PS1 for formatting and attribution requirements.
- Note that is is fine to hand-write solutions to the the non-coding questions, but make sure your writing is neat and insert any hand-written parts in order into your final submission.

1. Consider a censored regression problem. We assume a simple linear regression model, $Y_i \sim \mathcal{N}(\beta_0 + \beta_1 b_i, \sigma^2)$. Suppose we have an IID sample, but that for any observation such that $Y_i > \tau$, all we are told is that Y_i exceeded the threshold and not its actual value. In a given sample, n_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 censoring in the left tail) is in measuring pollutants, for which values below a threshold are reported as below the limit of detection. Another real world example is US tax revenue data where the incomes of wealthy taxpayers may be reported as simply exceeding, say, 1 million dollars.

   a. Design an EM algorithm to estimate the three parameters, $\theta = (\beta_0, \beta_1, \sigma^2)$, taking the complete data to be the available fully-reported data plus the actual values of the censored observations. You'll need to make use of $E(W|W > \tau)$ and $\text{Var}(W|W > \tau)$ where W 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. A few hints:

      i. Considering the notation we used in class when discussing EM, it's natural to think of Z as the (unobserved) values of the censored observations. You can think of $Z_i > \tau$ for the censored observations as being part of X (using X in the way that is is used in the class notes on EM), along with the uncensored observations.

      ii. When you write out the complete data log-likelihood, it's helpful to set it up so there is a term involving the observed values and a term involving the censored values. When you condition the latter term, you'll be conditioning on the censored values being bigger than threshold.

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

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

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

iv. You should recognize that your expected log-likelihood can be expressed as a regression of $\{Y_{obs}, m^t\}$ on $\{b\}$ where Y_{obs} are the non-censored data and $\{m_i^t\}$, $i = 1, \ldots, n_c$ 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 an additional term that you should interpret statistically.

v. You should be able to analytically maximize the expected log likelihood.

b. Propose reasonable starting values for the three parameters as functions of the observations.

c. Write an Python function, with auxiliary functions as needed, to estimate the parameters. Make use of the initialization from part (b). You may use `statsmodels` for computing β^{t+1}. You'll need to include criteria for deciding when to stop the optimization. Test your function using data simulated based on the code in `ps8.py` with (a) a modest proportion of exceedances expected, say 20%, and (b) a high proportion, say 80%.

2. 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, $P(Y_i > \tau; b_i, \beta_0, \beta_1, \sigma^2)$.

   a. Write a function that calculates the negative log-likelihood of the observed data.

   b. Estimate the parameters for your test cases using `scipy.optimize.minimize()` with the BFGS option. You will want to consider reparameterization. Compare how many iterations EM and BFGS take. Note that this provide a nice test of your EM derivation and code, since you should get the same results from the two optimization approaches. Calculate the estimated standard errors based on the inverse of the Hessian. Note that the `hess_inv` returned by `minimize` is probably NOT a good estimate of the Hessian as it seems to just be the approximation built up during the course of the BFGS iterations and not a good numerical derivative estimate at the optimum. Try using `numdifftools` and compare to what is seen in `hess_inv`.

   c. As part of this, try a variety of starting values and see if you can find ones that cause the optimization **not** to converge using BFGS. Does Nelder-Mead converge with those starting values?