STAT 578 – Advanced Bayesian Modeling – Fall 2018

Assignment 4

The groundbreaking 1929 paper[1] by Edwin Hubble offered evidence for expansion of the universe. Astronomical observations showed that "extra-galactic nebulae" (other galaxies) tended to be moving away at a rate roughly proportional to their distance:

$$v \;\approx\; H_0 D$$

where v is the radial velocity of the galaxy (away from us) in km/s, D is its proper distance in megaparsecs (Mpc), and H_0 is called the Hubble constant. The relationship is not exact – each galaxy also has its own "peculiar velocity" that is unrelated to the expansion.

File `hubbledata.txt` contains Hubble's original data on 24 astronomical objects, with their assumed distance and radial velocity.

(a) [2 pts] Plot the data points: radial velocity versus distance.

(b) Consider a normal-theory simple linear regression model of radial velocity on distance of the form

$$v_i \mid \beta, \sigma^2, D_i \;\sim\; \text{indep. N}\big(\beta_1 + \beta_2 D_i,\, \sigma^2\big) \qquad i = 1, \ldots, 24$$

Of course, the theory predicts that the intercept β_1 will be exactly zero, but your initial model will not assume this. Also, according to theory, the slope β_2 should be H_0. Use independent priors

$$\beta_1, \beta_2 \;\sim\; \text{iid N}\big(0,\, 10000^2\big)$$

$$\sigma^2 \;\sim\; \text{Inv-gamma}(0.0001,\, 0.0001)$$

Do not standardize or center any variables.

(i) [2 pts] List an appropriate JAGS model.

Now run your model. Make sure to use multiple chains with overdispersed starting points, check convergence, and monitor β_1, β_2, and σ^2 for at least 2000 iterations (per chain) after burn-in.

(ii) [2 pts] List the coda summary of your results for β_1, β_2, and σ^2.

(iii) [2 pts] Give the approximate posterior mean and 95% posterior credible interval for the slope. (Does H_0 appear to be positive?)

(iv) [2 pts] Give the approximate posterior mean and 95% posterior credible interval for the intercept. (Does your interval contain zero?)

(c) Consider the model of the previous part, but without the intercept (i.e., assuming the intercept is zero, as theory predicts). This is sometimes called *regression through the origin.* Use the same priors as before for the remaining parameters.

[1] Edwin Hubble, A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae, *Proceedings of the National Academy of Sciences*, vol. 15, no. 3, pp. 168–173, March 1929

(i) [2 pts] List your modified JAGS model.

Now run your model. Make sure to use multiple chains with overdispersed starting points, check convergence, and monitor parameters for at least 2000 iterations (per chain) after burn-in.

(ii) [2 pts] List the coda summary of your results for all parameters.

(iii) [2 pts] Give the approximate posterior mean and 95% posterior credible interval for the slope.

(iv) [2 pts] Compare the change in the posterior mean of the slope (versus part (b)) to its posterior standard deviation. (Has it changed very much relative to the standard deviation?) Also, is its credible interval wider or narrower than before?

(d) One way to check for evidence against the assumption that the intercept is zero is to produce a posterior predictive p-value based on the no-intercept model. Consider test quantity

$$T(y, X, \theta) \quad = \quad |\widehat{\mathrm{cor}}(\varepsilon, x_D)|$$

where $\widehat{\mathrm{cor}}(\varepsilon, x_D)$ is sample correlation between the error vector ε (*not* standardized) and the vector x_D of distances D in the data. The larger this quantity is for the no-intercept model, the less well that model fits the data (since, if a regression model actually fits, the errors should ideally be uncorrelated with the predictor).

Use your JAGS simulations from the previous part. (Suggestion: Apply `as.matrix` to the output of `coda.samples` to obtain a matrix of simulated parameter values.)

(i) [2 pts] Show R code for computing the simulated error vectors ε (as rows of a matrix).

(ii) [2 pts] Show R code for computing simulated *replicate* error vectors $\varepsilon^{\mathrm{rep}}$ (as rows of a matrix), which are the error vectors for the replicate response vectors y^{rep}.

(iii) [2 pts] Show R code for computing the simulated values of $T(y, X, \theta)$ and the simulated values of $T(y^{\mathrm{rep}}, X, \theta)$.

(iv) [2 pts] Plot the simulated values of $T(y^{\mathrm{rep}}, X, \theta)$ versus those of $T(y, X, \theta)$, with a reference line indicating where $T(y^{\mathrm{rep}}, X, \theta) = T(y, X, \theta)$.

(v) [2 pts] Compute the approximate posterior predictive p-value, and make an appropriate conclusion based on it. (Does it provide evidence that the no-intercept model does not fit?)

Remark: Modern determinations of H_0 vary around 70 (km/s)/Mpc, which is probably much different than what you obtained. Hubble's distance data was systematically in error because he had no accurate way to measure extra-galactic distances.

Total: 28 pts