STAT 221: Problem Set 2
Kevin Eskici and Willy Xiao
Due: Oct 7, 2014

Task 1: Examine the Posterior

Posterior is: $f(\mu, \sigma^2, \log \vec{\theta} | Y)$

$$f(\mu, \sigma^2, \log \vec{\theta} | Y) \propto f(Y | \log \vec{\theta}) * f(\log \vec{\theta} | \mu, \sigma^2) * f(\mu, \sigma^2)$$

$$\propto \prod_{j=1}^{J} \left[\left(\prod_{i=1}^{N} e^{-w_j e^{\log \theta_j}} \left(w_j e^{\log \theta_j} \right)^{Y_{ji}} \right) * \frac{1}{\sigma} e^{\frac{-(\log \theta_j - \mu)^2}{2\sigma^2}} * \frac{1}{\sigma^2} \right]$$

The Posterior of $\log \theta$ conditional on all the other parameters is the same as the previous equation, except we can drop the prior on μ and σ^2 because they're given:

$$f(\log \theta | Y, \mu, \sigma^2) \propto \prod_{j=1}^{J} \left[\left(\prod_{i=1}^{N} e^{-w_j e^{\log \theta_j}} \left(w_j e^{\log \theta_j} \right)^{Y_{ji}} \right) * \frac{1}{\sigma} e^{\frac{-(\log \theta_j - \mu)^2}{2\sigma^2}} \right]$$

To check the shape of the function, we can look at the second derivative of the log-posterior:

Let $\log p = \log f(\log \theta | Y, \mu, \sigma^2)$

$$= \sum_{j=1}^{J} \left(\sum_{i=1}^{N} \left(-w_j e^{\log \theta_j} + Y_{ji} * \left(\log w_j + \log e^{\log \theta_j} \right) \right) - \log \sigma - \frac{(\log \theta_j - \mu)^2}{2\sigma^2} \right)$$

$$\frac{\partial \log p}{\partial \log \theta_j} = \sum_{j=1}^{J} \left(\sum_{i=1}^{N} \left(-w_j e^{\log \theta_j} + Y_{ji} \log \theta_j \right) - \frac{\log \theta_j - \mu}{\sigma^2} \right)$$

$$\frac{\partial^2 \log p}{\partial \log \theta_j^2} = \sum_{j=1}^{J} \left(\sum_{i=1}^{N} \left(-w_j e^{\log \theta_j} + Y_{ji} \right) - \frac{1}{\sigma^2} \right)$$

Because the second derivative of the log-posterior is monotonically decreasing with respect to $\log \vec{\theta}$, that means our function is unimodal (ie there's a single peak).

Task 2: Write functions to simulate data from the model

Check keskici_wxiao_ps2_functions.R

Task 3: Evaluate coverage for a simple case

First: Estimate the amount of simulations we can do:

a) The time it took to run one simulation on my Macbook Air was roughly 23 seconds. Accounting for startup costs (copying to multiple machines, however Odyssey decides to manage nodes etc.,), we roughly said that each MCMC simulation would take 25 seconds.

b) Calculating, we have

$$(60 \text{ seconds}) \times (60 \text{ minutes}) \times (12 \text{ nodes}) = 43200 \text{ seconds of runtime.}$$

Then

$$\frac{43200}{(4 \text{ parameters})}/(25 \text{ seconds per simulation}) = 432 \text{ simulations per parameter.}$$

We ultimately decided to go with 360 simulations per parameter to guarantee that we don't go over the hard one-hour time limit. Also 360 is a number that is divisible by 12 and easily modeled on our local machines to test (e.g. running 36 simulations rather than 360), which makes writing the .slurm job a little bit easier.

Second: Decide how many theta's to draw and how many Y's to draw:
Ultimately we weren't sure how best to do this. The example on the pset had more theta draws than Y draws, so we decided to do the same.

$$\text{theta.draws} = 30$$
$$\text{Y.draws} = 12$$

Third: RUN! :)

Unfortunately, the timing was wildly off. It didn't take into account that we had to calculate the coverage after each theta draw which means we had to recalculate how many times simulations we could do! We tried 12×15 (ie half of what it was before, and it still went too long), 12×12 was close. Ultimately we ran it for:

$$\text{theta.draws} = 9$$
$$\text{Y.draws} = 15$$

Task 4: Evaluate coverage with exposure weights

We ran the same simulation as Task 3, except we added in the exposure weights. Results discussion below, code in relevant file.

Task 5: Evaluate coverage with model misspecifcation:

We ran the same simulation as Task 4, substituting in rASL instead of rnorm as our function to generate $\log \theta_j$. For running the MCMC simulation, we left the parameters mu and sigmasq to be their defaults 0 and 1 respectively.

Task 6: INSERT SHIT HERE

Appendix:

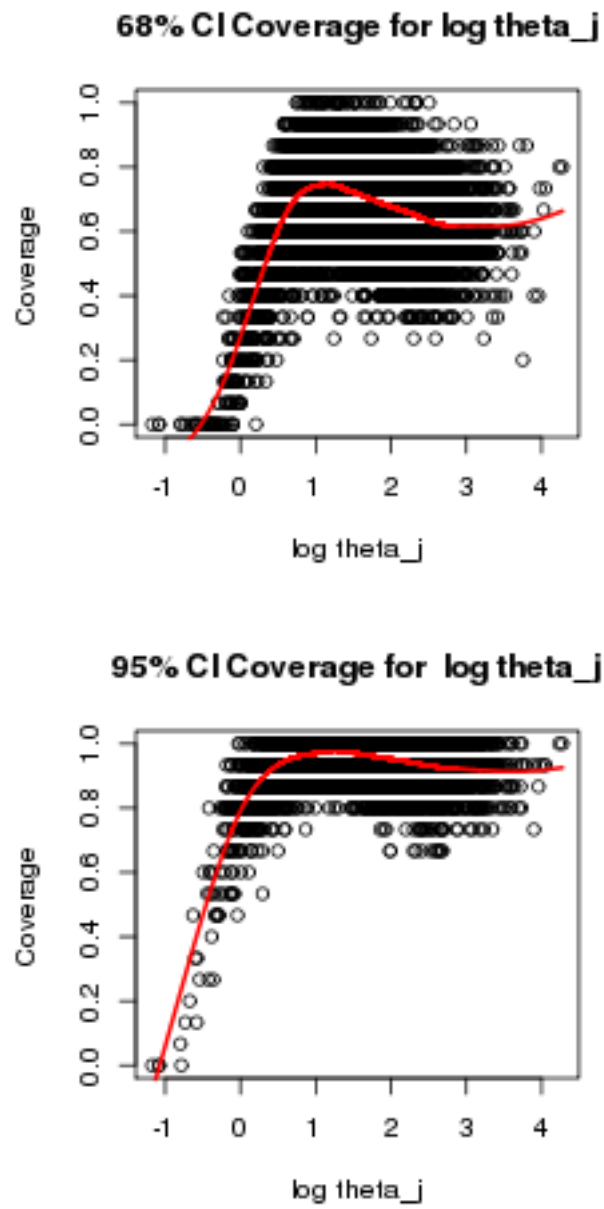Figure 1: Task 3 coverage plots for $log(\theta_j)$'s : $\mu = 1.6$, $\sigma^2 = 0.7^2$

68% CI Coverage for log theta_j

95% CI Coverage for log theta_j

Figure 2: Task 3 coverage plots for $log(\theta_j)$'s : $\mu = 2.5$, $\sigma^2 = 1.3^2$

68% CI Coverage for log theta_j

95% CI Coverage for log theta_j
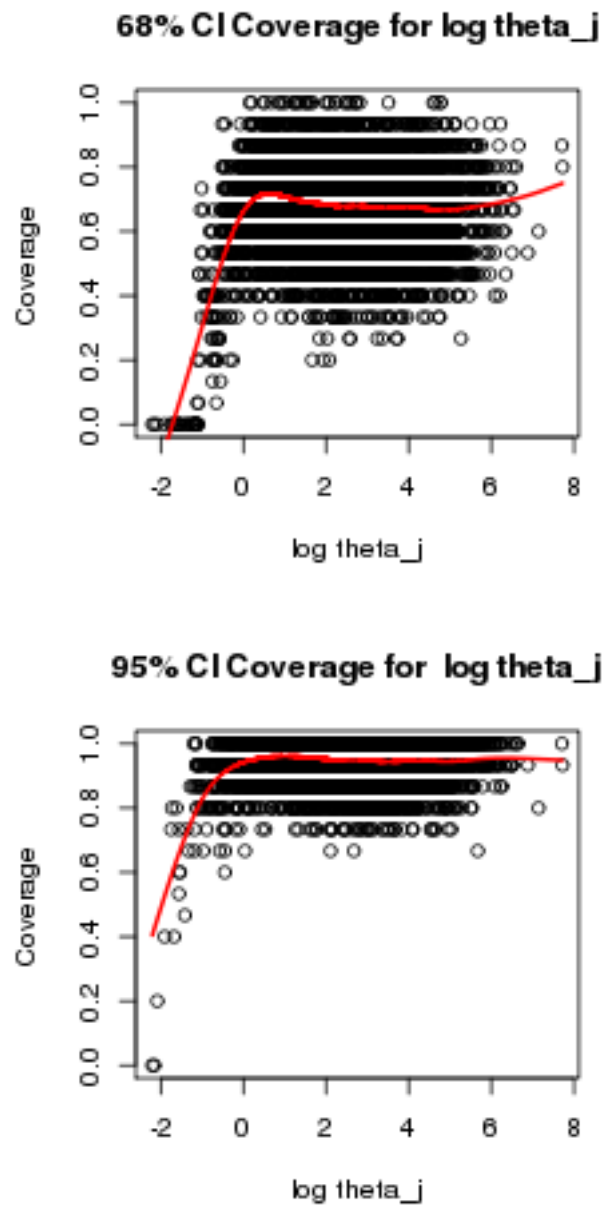
Figure 3: Task 3 coverage plots for $log(\theta_j)$'s : $\mu = 5.2$, $\sigma^2 = 1.3^2$

68% CI Coverage for log theta_j

95% CI Coverage for log theta_j
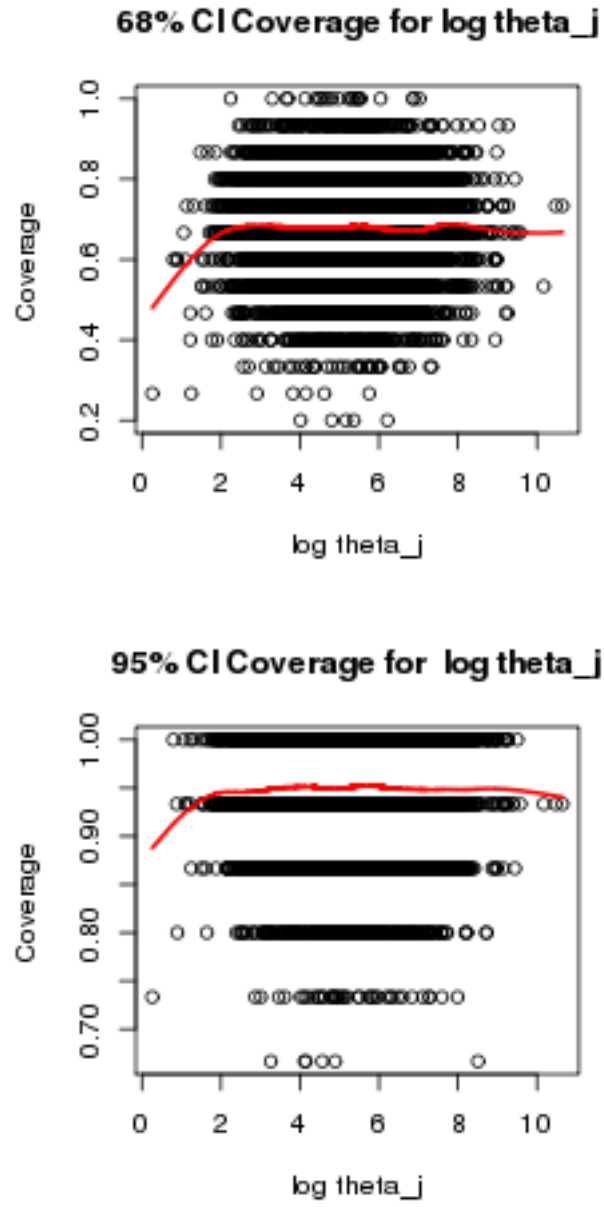
Figure 4: Task 3 coverage plots for $log(\theta_j)$'s : $\mu = 4.9$, $\sigma^2 = 1.6^2$

68% CI Coverage for log theta_j

95% CI Coverage for log theta_j
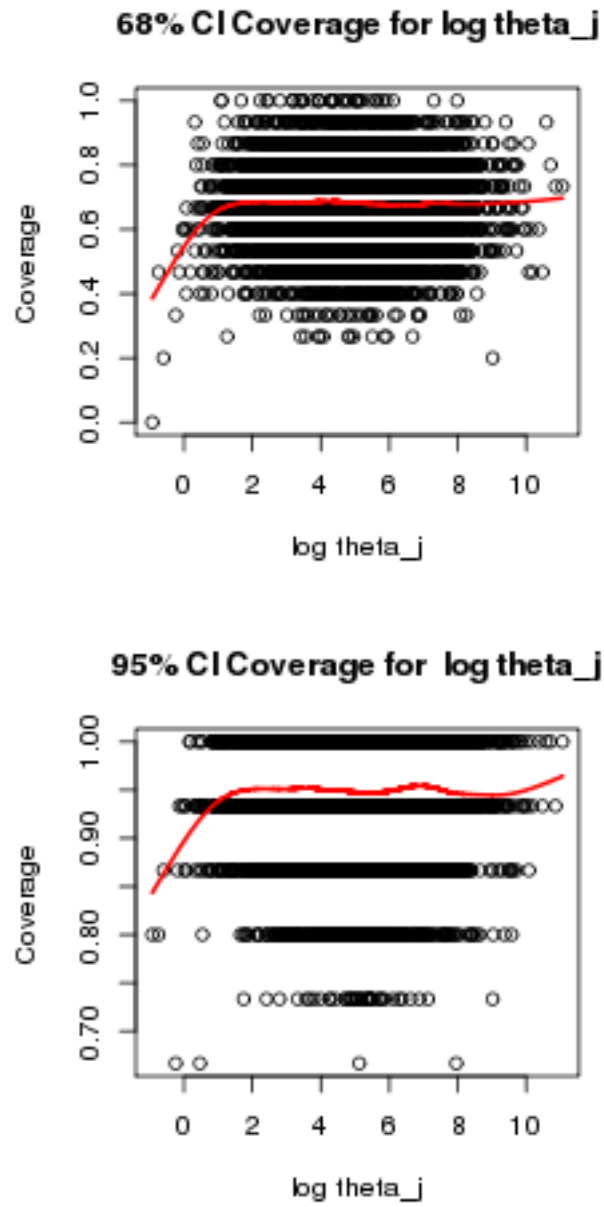
Figure 5: Task 4 coverage plots for $\mu = 1.6$, $\sigma^2 = 0.7^2$
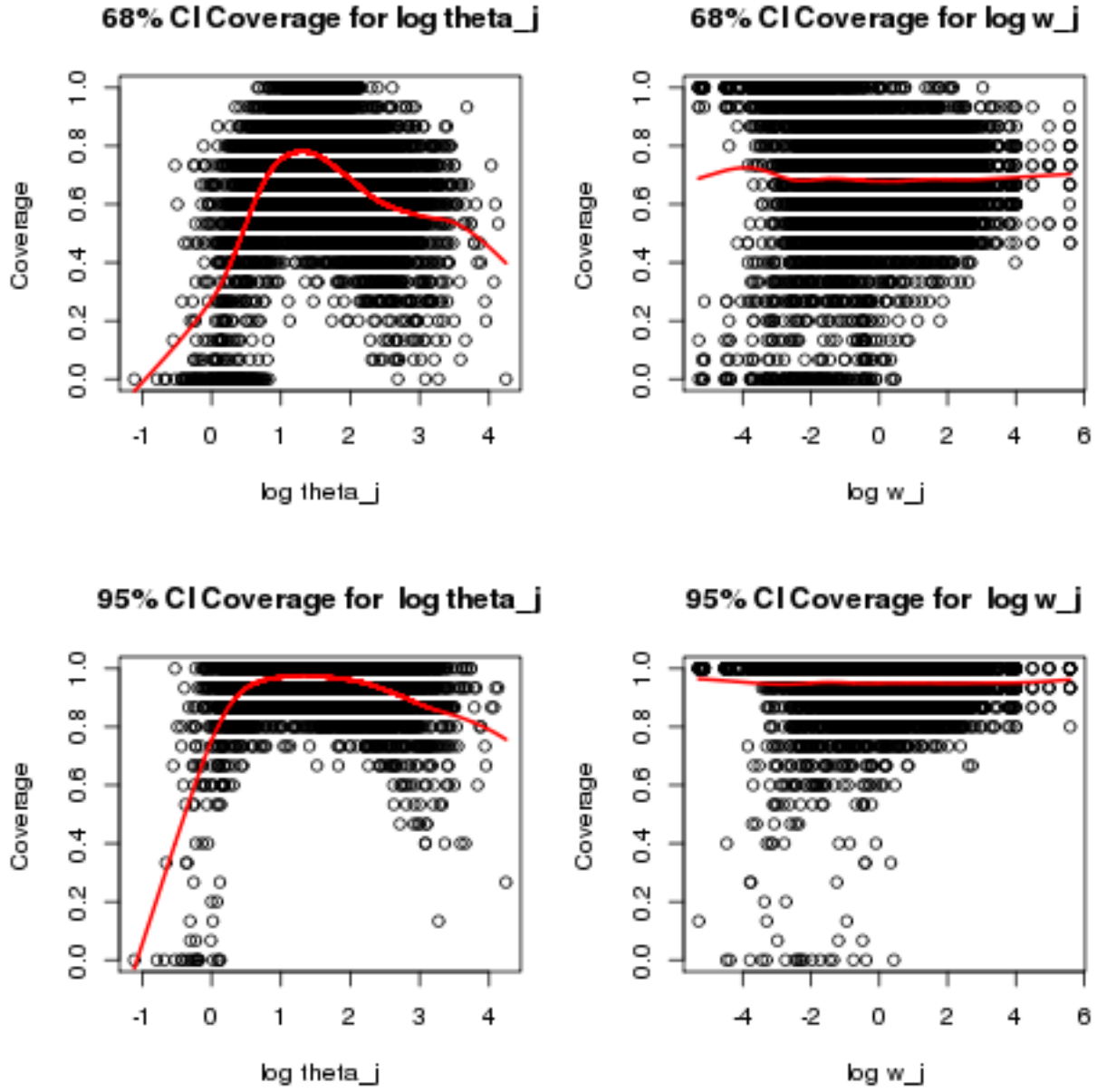
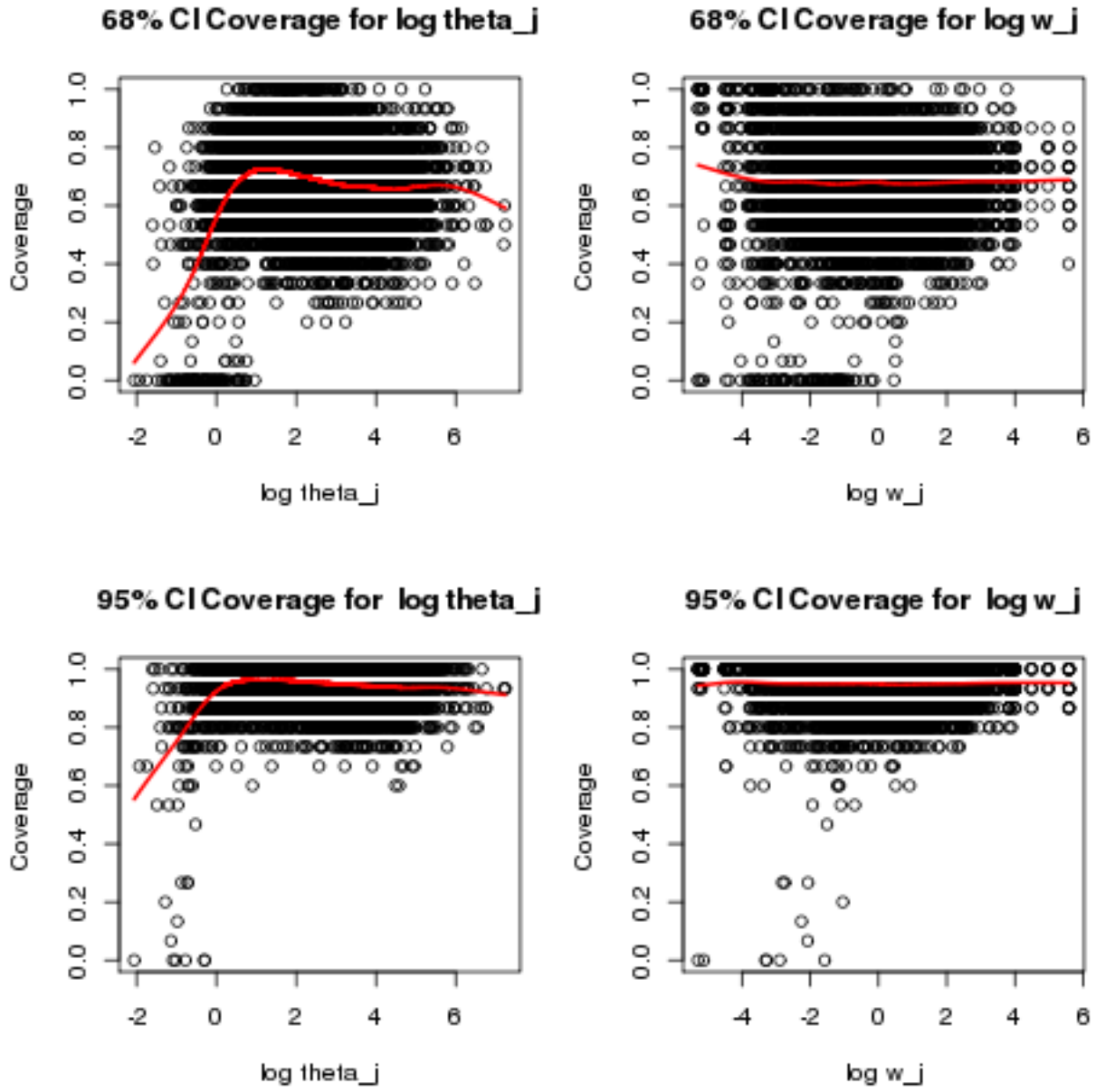Figure 6: Task 4 coverage plots for $\mu = 2.5$, $\sigma^2 = 1.3^2$

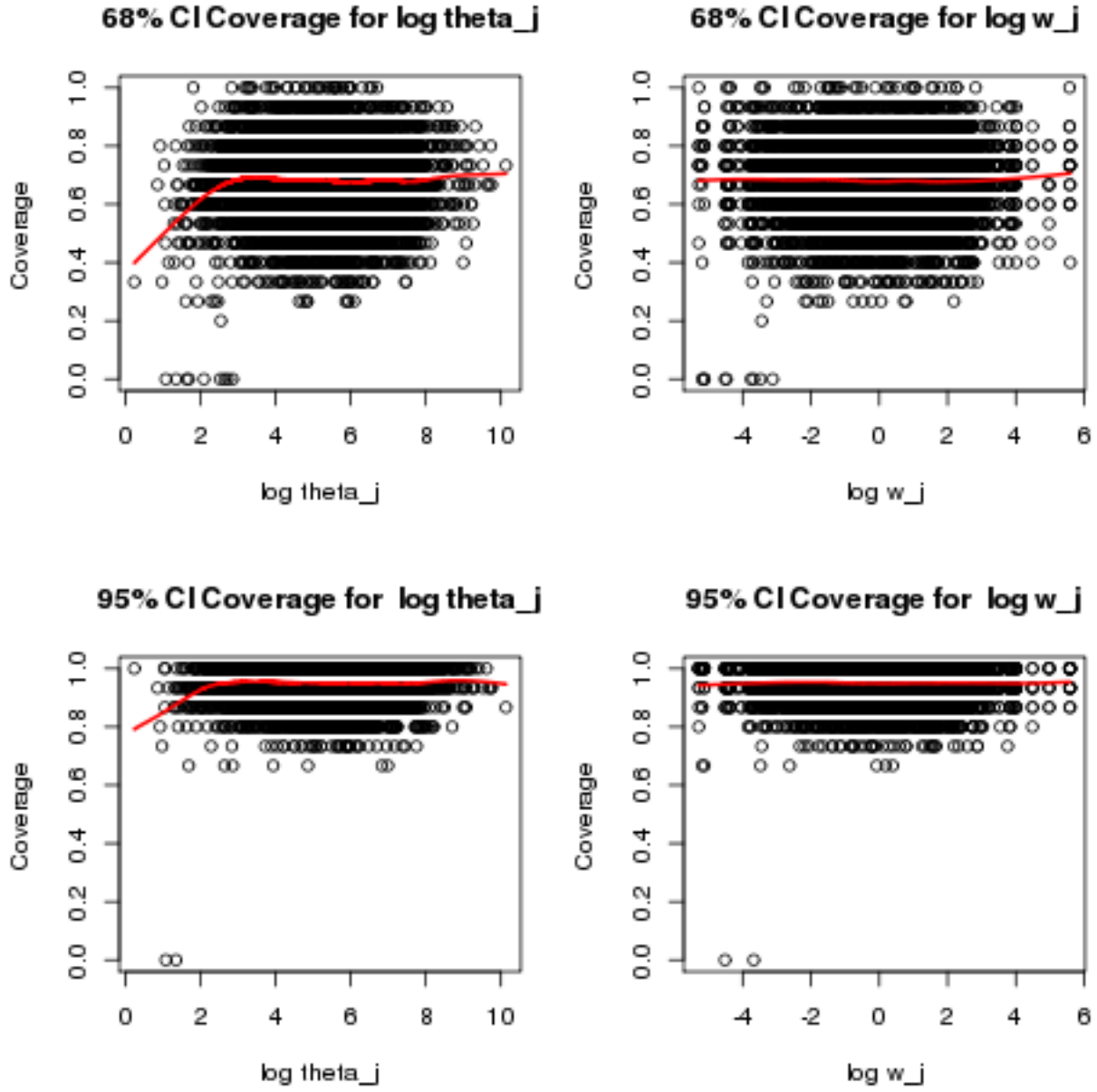Figure 7: Task 4 coverage plots for $\mu = 5.2$, $\sigma^2 = 1.3^2$

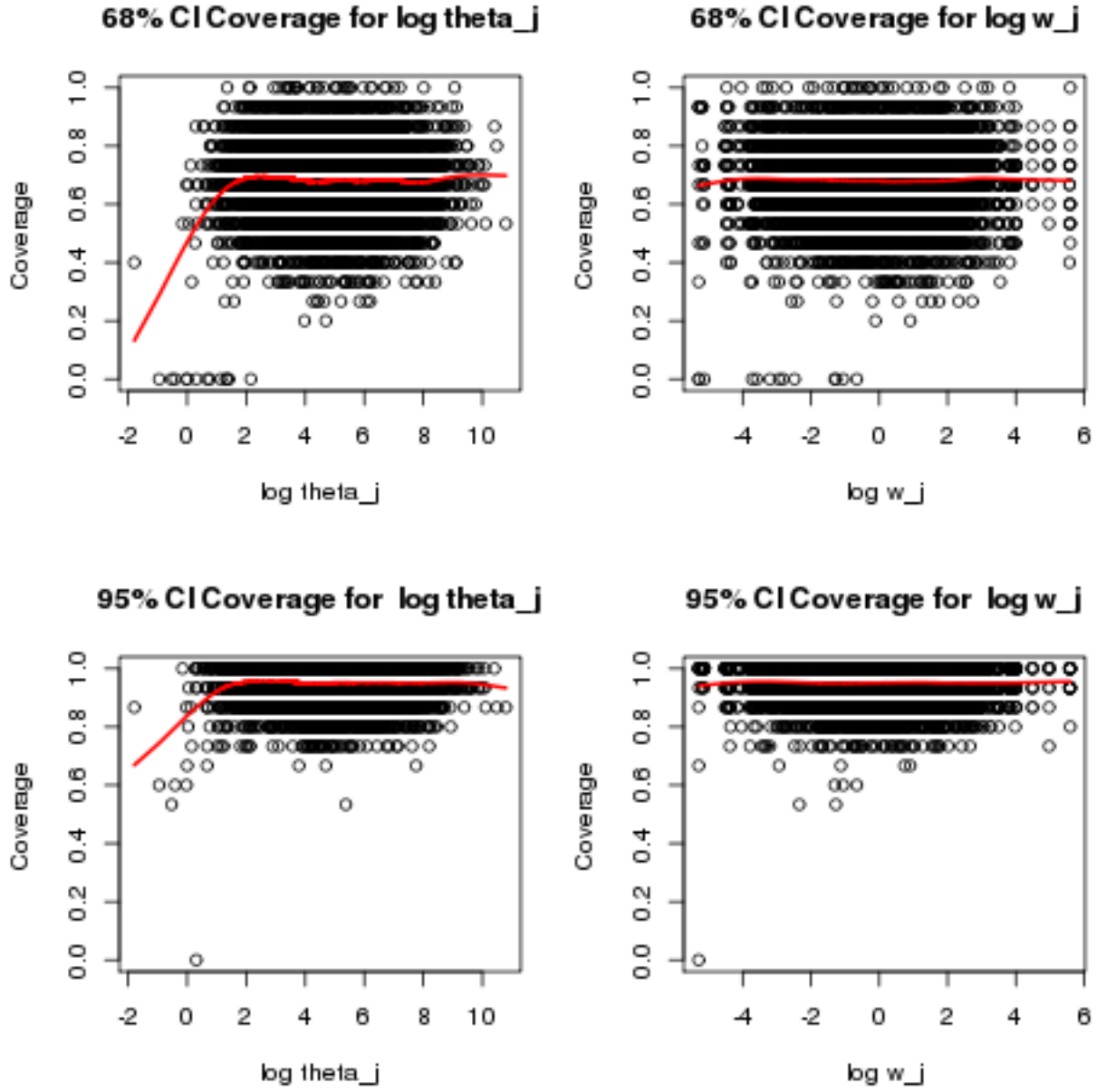Figure 8: Task 4 coverage plots for $\mu = 4.9$, $\sigma^2 = 1.6^2$

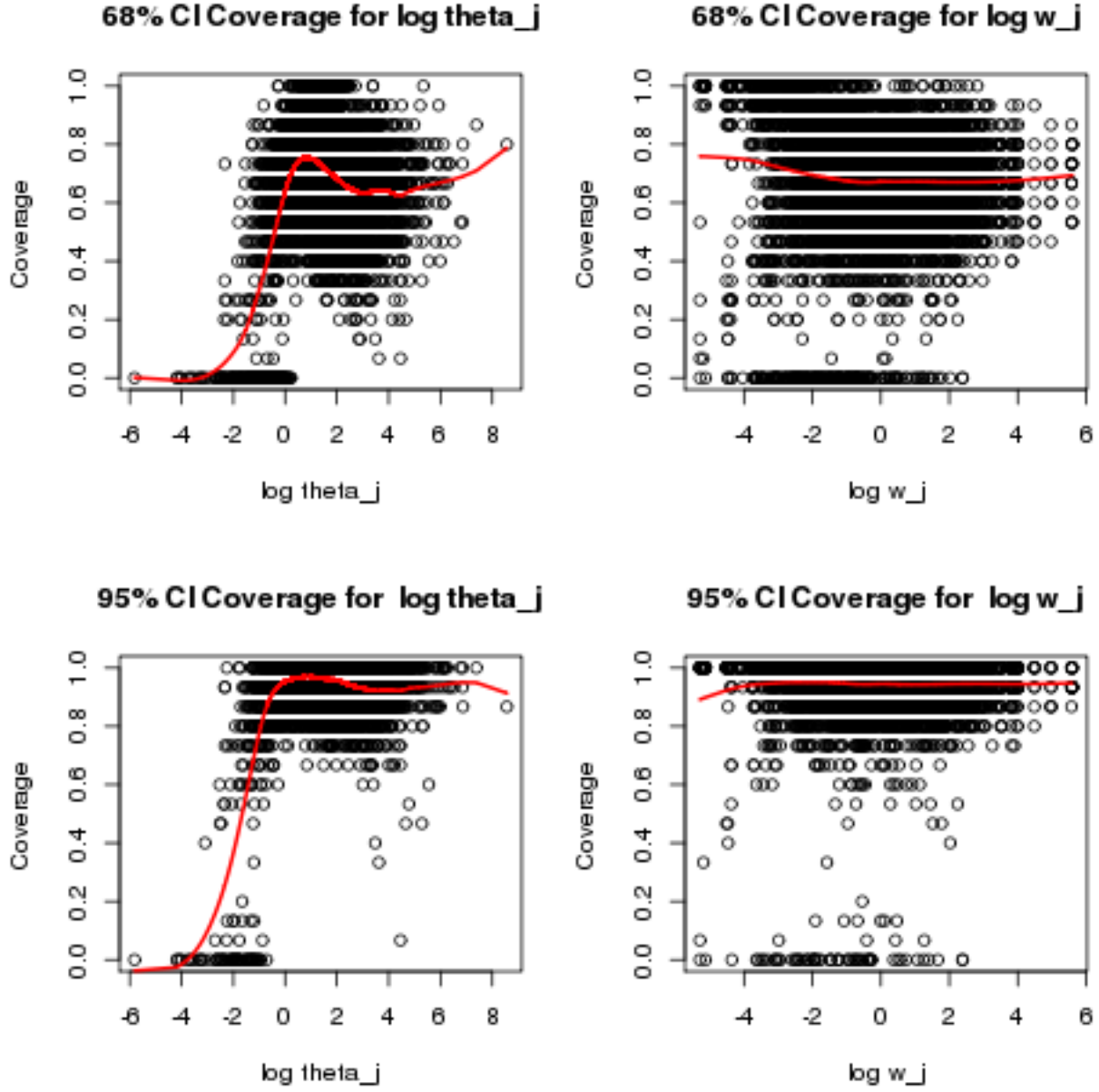Figure 9: Task 5 coverage plots for $log(\theta_j)$'s : $x_0 = 1.6$, $m = 0, b = 1.3$

68% CI Coverage for log theta_j

68% CI Coverage for log w_j

95% CI Coverage for log theta_j

95% CI Coverage for log w_j

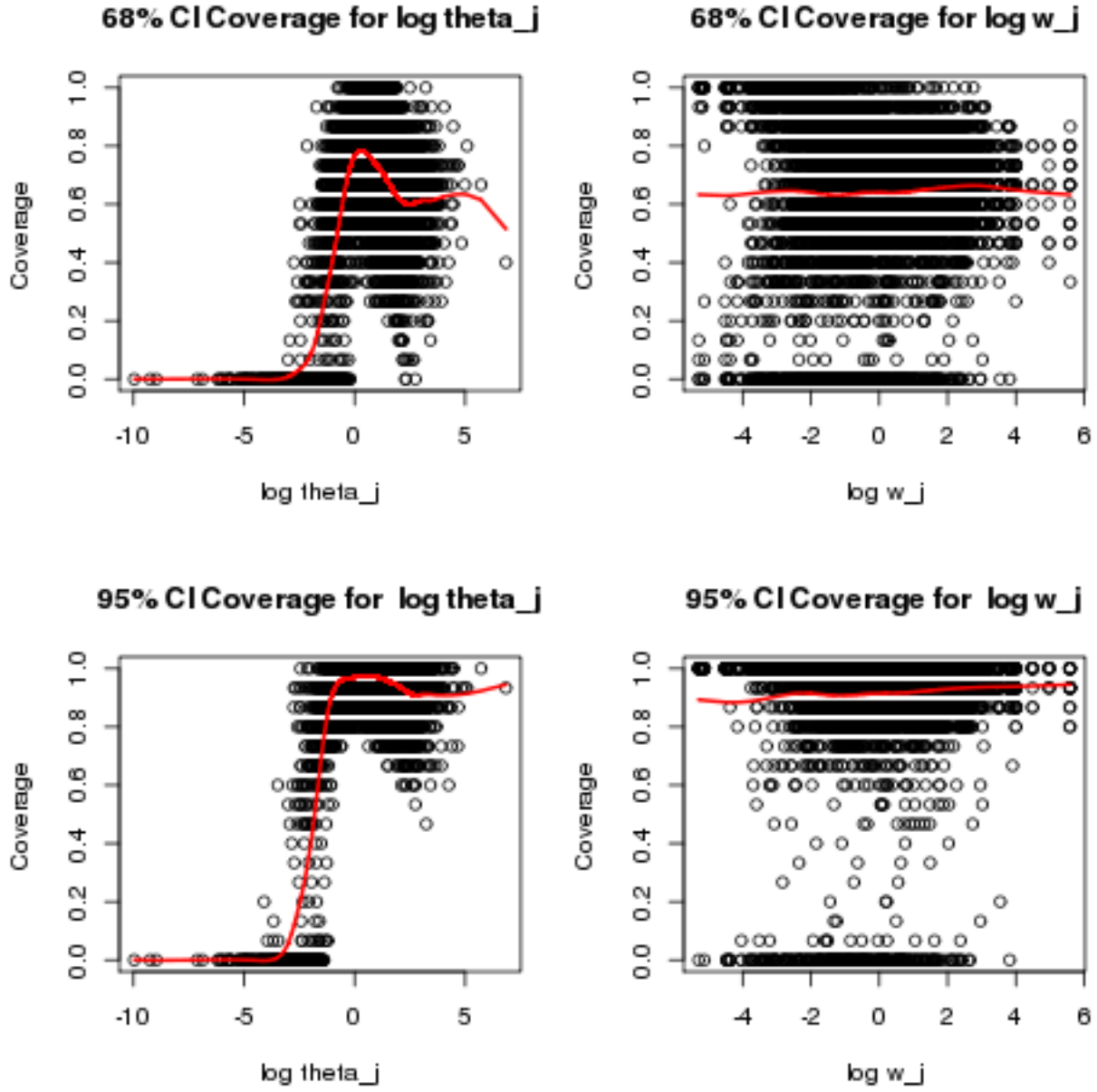Figure 10: Task 5 coverage plots for $log(\theta_j)$'s : $x_0 = 1.6$, $m = -0.7$, $b = 1.3$

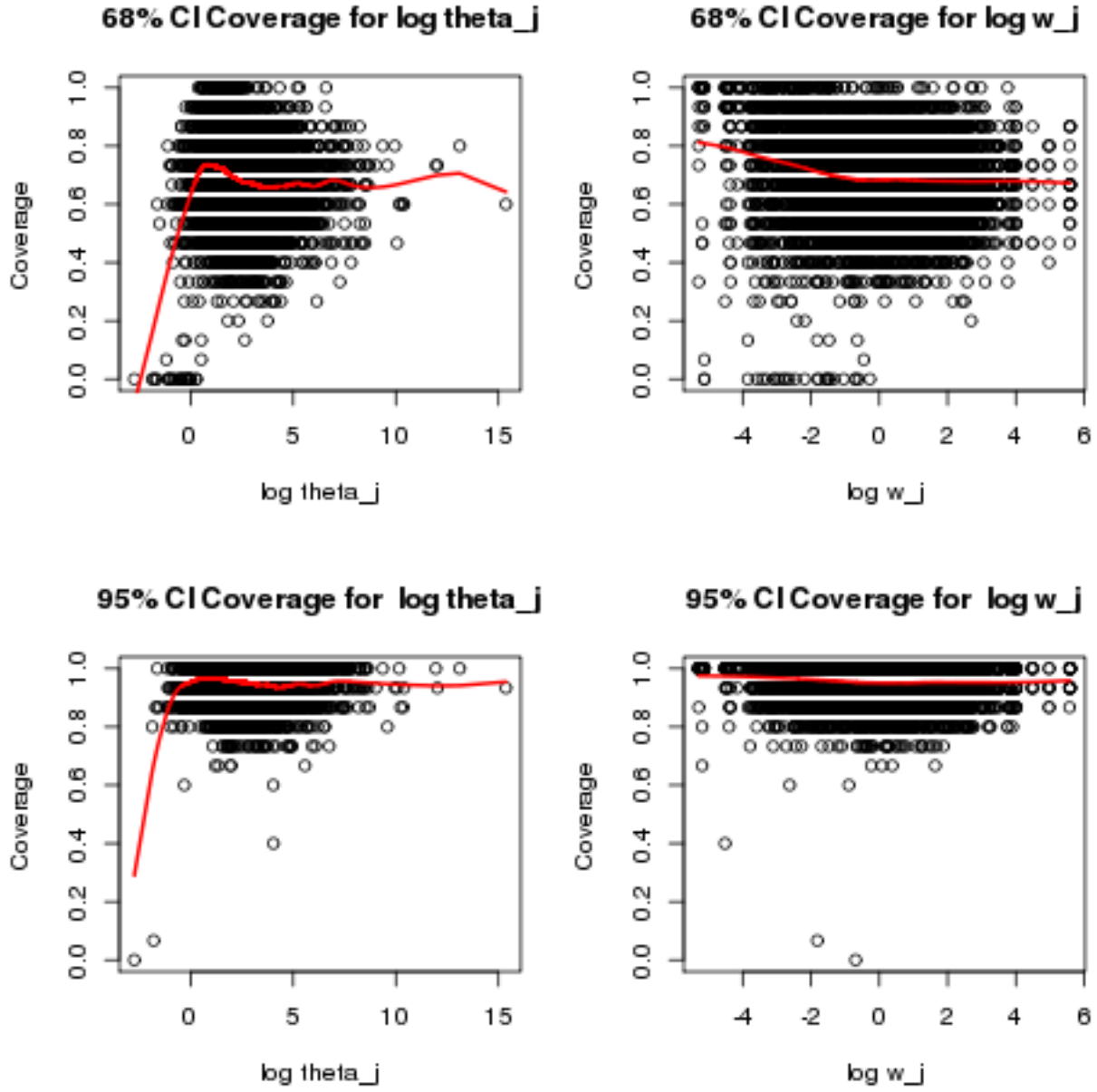Figure 11: Task 5 coverage plots for $log(\theta_j)$'s : $x_0 = 1.6$, $m = -.7, b = 1.3$

Figure 12: Task 5 coverage plots for $x_0 = 1.6$, $m = 0, b = 2.6$