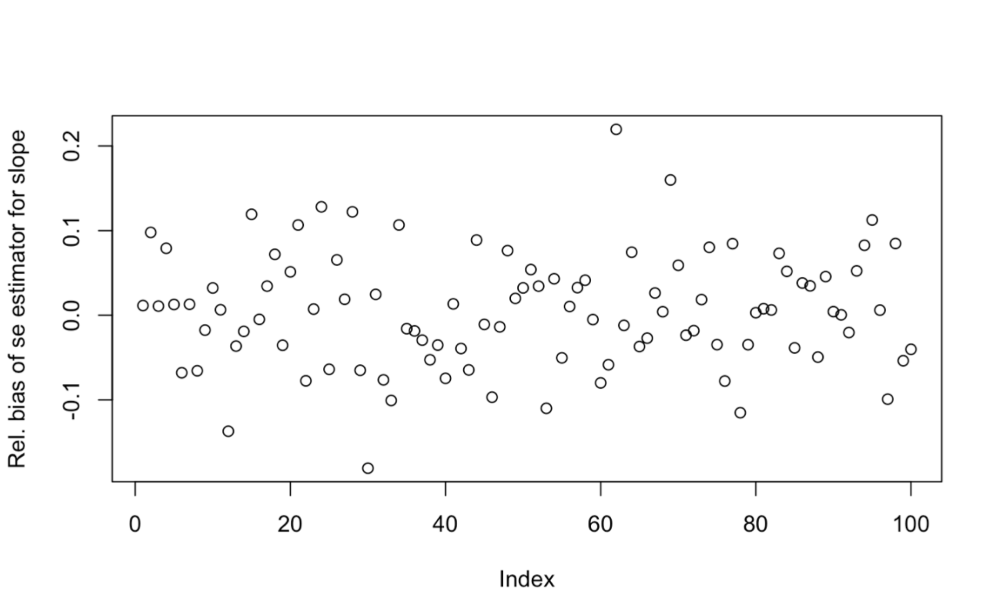
Did simulations with a simplified setting, for $i = 1, \ldots, 100$ (sample size: 5,000):

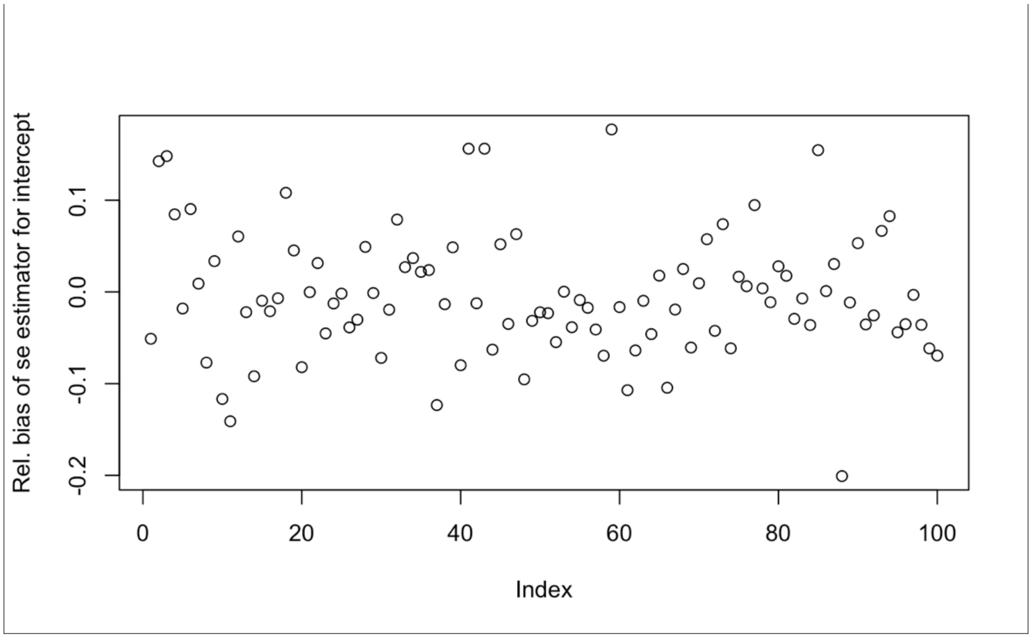
- X ~ N(0, 1); Logit(P(Y=1)) = 1 + 2 * X
- Fit the model using X and Y and we call this model m_i.
- Repeat the following procedure for 100 times:
    - External X ~ N(0,1); Logit(P(External Y=1)) = 1 + 2 * X
    - Obtain the estimates of calibration slope and intercept
- Take the average of the estimates and use the average values as the true value of model calibration slope and intercept
- Repeat the following procedure for 100 times:
    - External X ~ N(0,1); Logit(P(External Y=1)) = 1 + 2 * X
    - Obtain the estimates of calibration slope and intercept and their standard error
    - Construct 95% CIs based on the point and standard error estimates and check if the true value falls in the CIs
- Take the average of the standard error estimates and calculate the sample standard deviation of the point estimates; calculate % of times when the true value falls in the CIs as the empirical coverage probability

Now we have fit 100 different models and for each model, we have a formula-based standard error estimate, an empirical standard error, and an empirical coverage probability for calibration slope and intercept, respectively

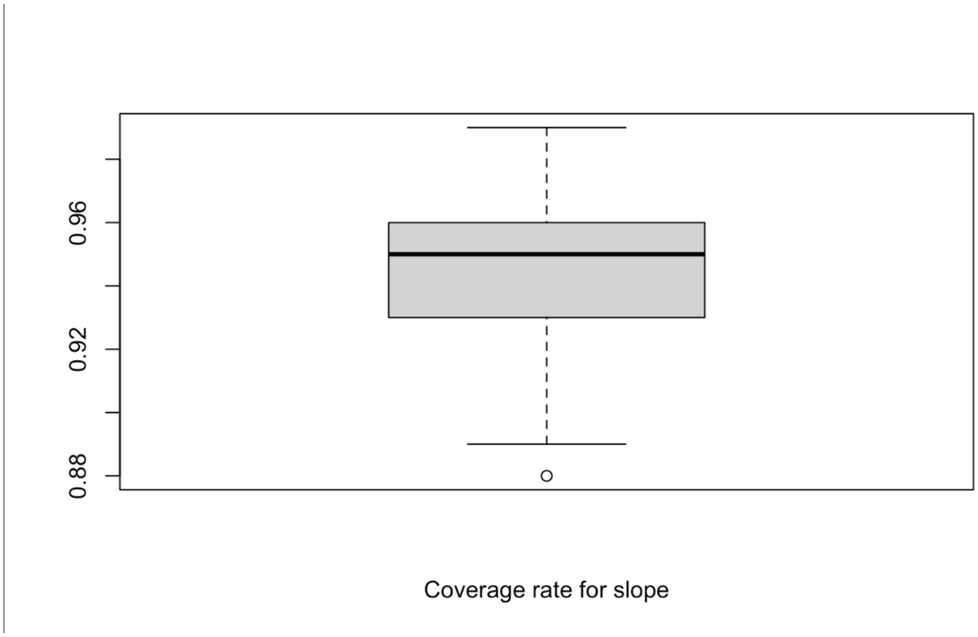
Relative difference between formula-based standard error and empirical standard error for calibration slope:



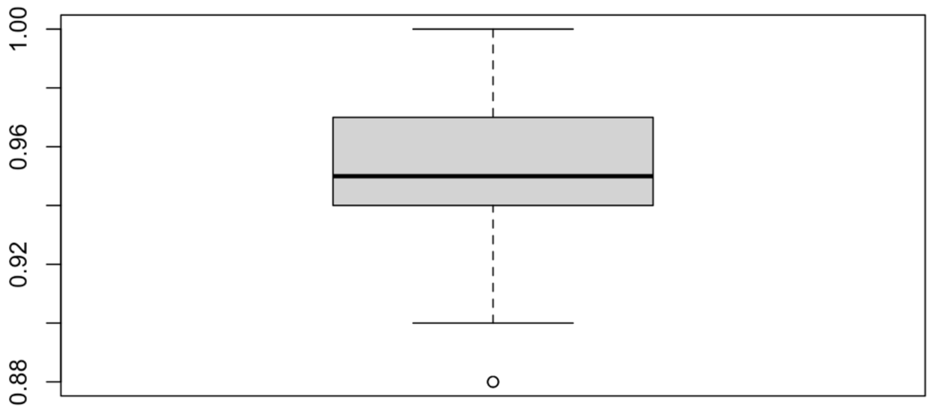
Relative difference between formula-based standard error and empirical standard error for calibration intercept:



Box-plot of coverage probability for calibration slope (mean value 0.9447):



Box-plot of coverage probability for calibration intercept (mean value 0.9505):



Coverage rate for intercept