Comparing the Accuracy of Three Predictive Information Criteria for Bayesian Linear Multilevel Model Selection

Sean Devine, Carl F. Falk, Ken A. Fujimoto

McGill University

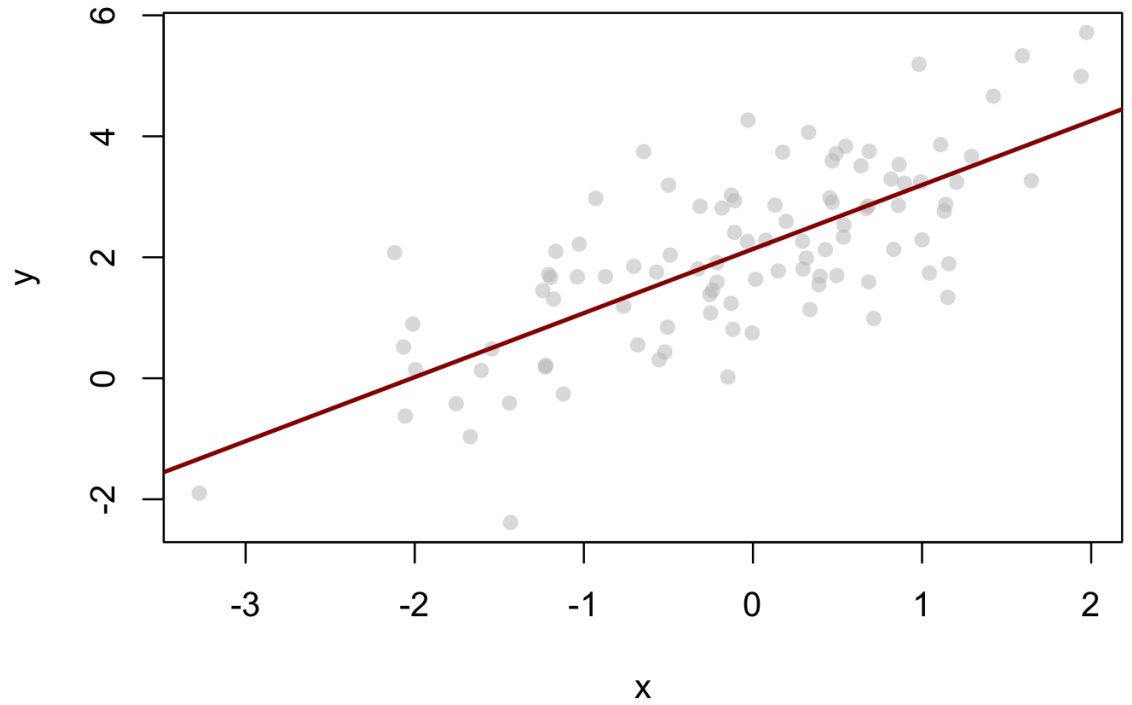
MMM 2024, UConn

[seandamiandevine@gmail.com]

Linear modeling

$$y_i = \beta_0 + \beta_1 X1_i + \epsilon$$

$$\epsilon = N(0, \sigma^2)$$



Linear **multilevel** modeling

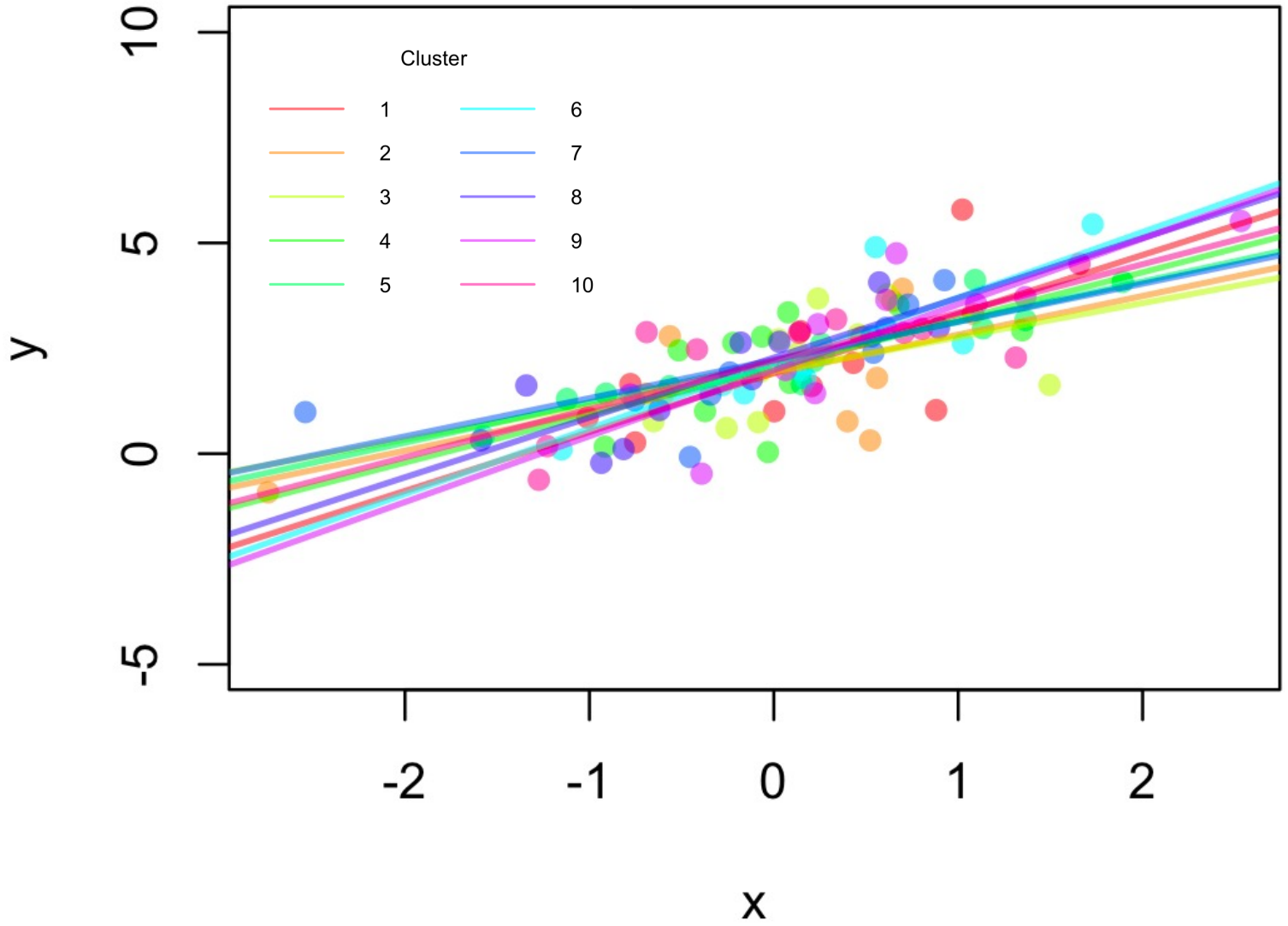
$$y_{ij} = \gamma_{00} + u_{0j} + \left(\gamma_{10} + u_{1j}\right)X1_{ij} + \epsilon$$

$$\left(u_{0j}, u_{1j}\right) = MVN(0, \Sigma)$$

$$\Sigma = \begin{bmatrix} \tau_{00} & \rho_{10} \\ \rho_{10} & \tau_{10} \end{bmatrix}$$

$$\epsilon = N(0, \sigma^2)$$

Linear **multilevel** modeling



The popularity of multilevel modeling

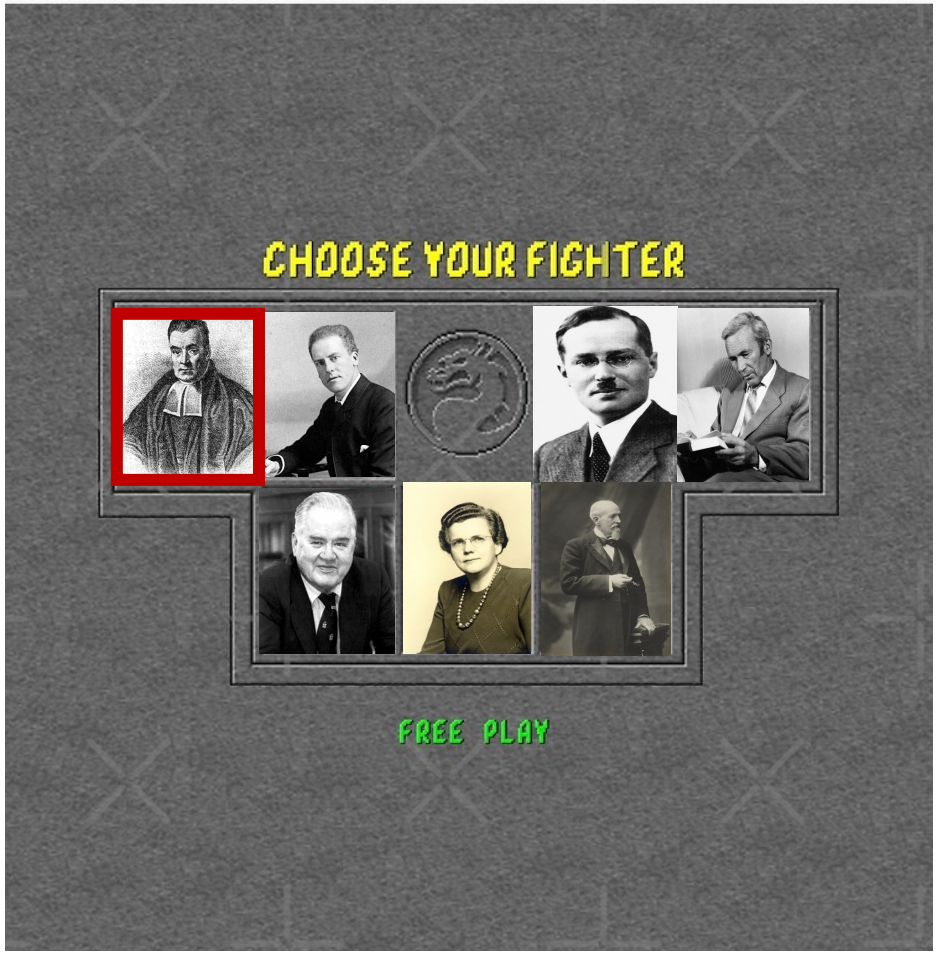
- From 2007 to 2017 there has been a threefold increase in the number of published psychology articles that utilize multilevel modeling (Huang, 2018)

- The advance of easily accessible software for fitting linear multilevel models has likely contributed to this trend (e.g., *lme4* in R, Bates et al., 2015)
  - Many such tools estimate multilevel models in a frequentist framework, using maximum likelihood estimation

The popularity of **Bayesian** multilevel modeling

- Similarly, advances in computational efficiency have contributed to the popularity of Bayesian methods for estimating multilevel models
  - Supported by software: *brms* (Bürkner, 2017), built on Stan (Carpenter et al., 2017)

- Bayesian estimation offers numerous advantages over frequentist analysis
  - Direct examination of posterior uncertainty
  - Incorporation of prior beliefs into parameter estimates



Model selection for Bayesian MLMs

- Despite some advantages of Bayesian estimation, multilevel **model selection** can be more complex
  - No simple significance test available (e.g., likelihood ratio test)

- This is problematic
  - It is often challenging to determine which of a set of candidate models is the “best model”
  - This question is of central importance because it is regularly this winning model from which scientific conclusions will be drawn

Information Criteria for Bayesian models

- Researchers have proposed several single-valued metrics that quantify the predictive accuracy of a Bayesian multilevel model: **Information Criteria**
  - Quantify the degree to which the observed data are likely to occur under the proposed model (while accounting for uncertainty in this likelihood)
  - Penalize more complex models in favor of more parsimonious models

- Three particularly popular metrics for Bayesian MLMs:
  - Deviance information criterion (DIC; Spiegelhalter et al., 2002)
  - Widely applicable information criterion (WAIC; Watanabe, 2010)
  - An approximation to the leave-one-out cross-validation based on Pareto smoothing of the importance sampling weights (LOO-CV; Vehtari et al., 2017)

DIC, WAIC, and LOO-CV (quickly)

DIC	WAIC	LOO-CV
$\text{DIC} = -2\ln p\left(\boldsymbol{y}\middle\vert\widehat{\widehat{\boldsymbol{\theta}}}\right) + 2p_D$ $p_D = -2\left(\frac{1}{S}\sum_{s=1}^{S}\ln p(\boldsymbol{y}\vert\boldsymbol{\theta}^s) - \ln p\left(\boldsymbol{y}\middle\vert\widehat{\widehat{\boldsymbol{\theta}}}\right)\right)$	$lppd = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\ln\frac{1}{S}\sum_{s=1}^{S}p(y_{ij}\vert\boldsymbol{\theta}^s)$ $p_W = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(V_{s=1}^{S}\ln p(y_{ij}\mid\boldsymbol{\theta}^s)\right)$ $\text{WAIC} = -2(lppd - p_W)$	$elppd = \sum_{j=1}^{J}\sum_{i=1}^{n_j}\ln\left(\frac{\sum_{s=1}^{S}w_{ij}^s p\left(y_{ij}\middle\vert\boldsymbol{\theta}^s\right)}{\sum_{s=1}^{S}w_{ij}^s}\right)$ $\text{LOO-CV} = -2elppd$

- More details provided in the supplement, but equations on the screen for your viewing pleasure
- To remember:
  - The output for each criterion is a **single float value** (e.g., 741.44) and this number differs between metrics; lower is better

Two other wrinkles to consider

Two other wrinkles to consider

- Normally, the relevant background would be covered at this point
  - Simulate data, fit Bayesian MLMs, compute ICs, see what happens

- However, two additional concepts are relevant here:
  1. Model selection uncertainty
  2. Marginal versus conditional parameter estimation

Wrinkle #1: Model selection uncertainty

- Historically, model selection has been based on singular point estimates of fit indices
  - Fit a series of candidate models, compute ICs, choose model with best IC
  - This approach ignores the sampling variability in selection criteria themselves (Preacher & Merkle, 2012)

- Accordingly, here we investigate MLM model selection using the "lowest value wins" approach and one which considers uncertainty:
  - DIC: 4 points distance (Spiegelhalter et al., 2002)
  - WAIC and LOO-CV: 1 standard errors in scores (Vehtari et al., 2017)

Wrinkle #2: Estimation methods

- In a multilevel modeling context, there is more than one possible form to the likelihood
- In the current context, there is debate as to whether using **conditional** or **marginal** estimation procedures yield more stable results (Merkle et al., 2019)

	Conditional	**Marginal**
Summary	Estimate cluster-level effects separately, under assumptions of multivariate normality	Estimate fixed effects and random variance coefficients, and infer, but not estimate, random effects

- Conditional is the default in *brms,* Marginal is used in *lme4*
- We implemented our own marginal estimation software in Stan to compare IC performance across estimation methods

This study

General approach

1. Simulate 28,800 datasets from a known multilevel linear model
2. The generated datasets varied with respects to
   - level-1 sample size (e.g., number of trials in an experiment)
   - level-2 sample size (e.g., number of participants in an experiment)
   - the magnitude of the random effect (larger or smaller random effect variance)
   - the magnitude of the fixed effect size (larger or smaller effect sizes)
3. We then modelled data using five specifications: the data generating model and five mispecified models
   - Using conditional or marginal estimation
4. We then computed ICs and selected models using a “lowest value wins” or model-selection uncertainty rule, described earlier
   - Computed accuracy as the rate of correctly identifying the data generating model

Data-generating model

Model A

$$y_{ij} = \gamma_{00} + u_{0j} + (\gamma_{10} + u_{1j})X_{1_{ij}} + \gamma_{20}X_{2_{ij}} + R_{ij}$$

- y_{ij} is the observed value for a continuous variable for observation i in cluster j
- γ_{00} is the fixed intercept (i.e., the mean value across clusters and observations)
- u_{0j} is the cluster-level deviation (i.e., random effect) from γ_{00} for the intercept
- γ_{10} is the fixed slope for X_1
- u_{1j} is the cluster-level deviation (random effect) from γ_{10}
- γ_{20} is the fixed slope for X_2
- $(u_{0j}, u_{1j}) \sim MVN(\mathbf{0}, \Sigma)$, where Σ is a variance–covariance matrix, $\Sigma = \begin{bmatrix} \tau_0^2 & \rho_{01} \\ \rho_{01} & \tau_1^2 \end{bmatrix}$
- $R_{ij} \sim N(0, \sigma^2)$

Other model specifications

Model	Form
A	$\boldsymbol{y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + \gamma_{20}X_{2_{ij}} + R_{ij}}$
B	$y_{ij} = \gamma_{00} + U_{0j} + \boldsymbol{\gamma_{10}X_{1_{ij}}} + \gamma_{20}X_{2_{ij}} + R_{ij}$
C	$y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + \boldsymbol{(\gamma_{20} + U_{2j})X_{2_{ij}}} + R_{ij}$
D	$y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + R_{ij}$
E	$y_{ij} = \gamma_{00} + U_{0j} + (\gamma_{10} + U_{1j})X_{1_{ij}} + \gamma_{20}X_{2_{ij}} + \boldsymbol{\gamma_{30}X_{3_{ij}}} + R_{ij}$

B. The (incorrect) absence of a random effect (wrongly assumes $\tau_1^2 = 0$)

C. The (incorrect) presence of a random effect (wrongly assumes $\tau_2^2 \neq 0$)

D. The (incorrect) absence of a fixed effect (wrongly assumes $\gamma_{20} = 0$)

E. The (incorrect) presence of a fixed effect (wrongly assumes $\gamma_{30} \neq 0$)

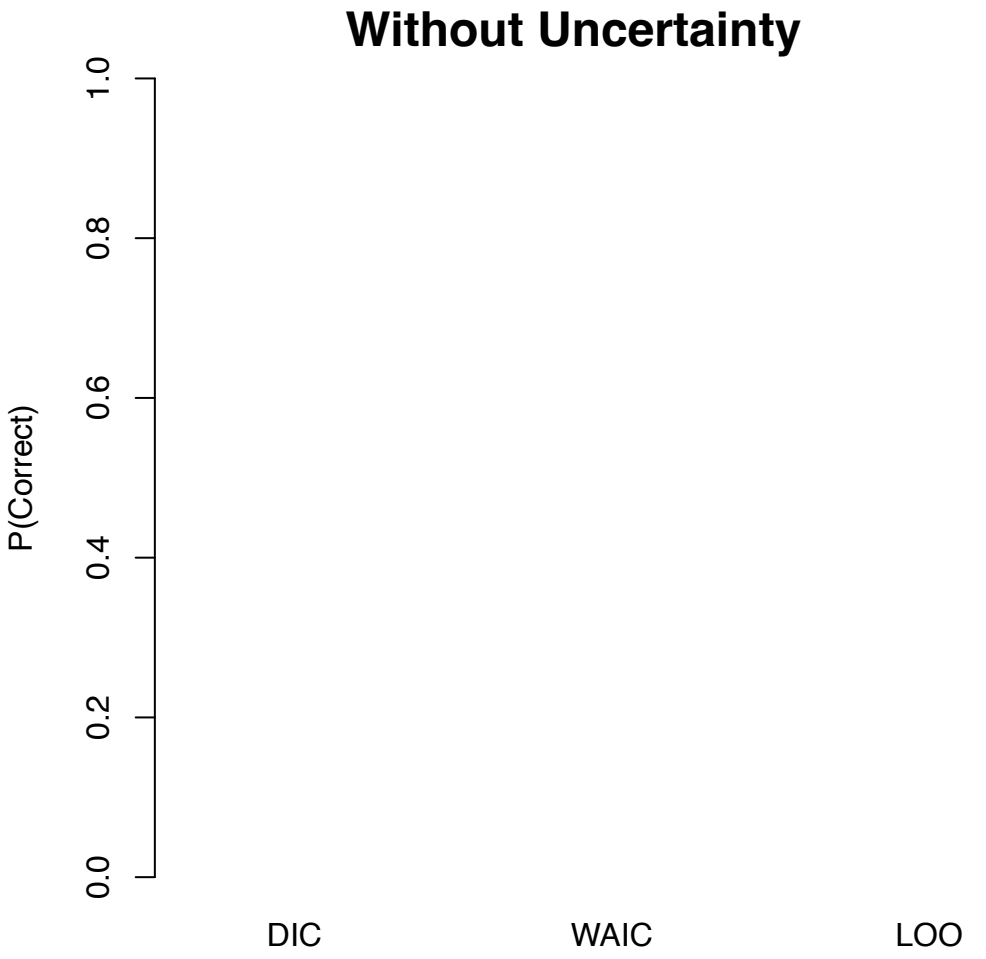
Simulation design matrix

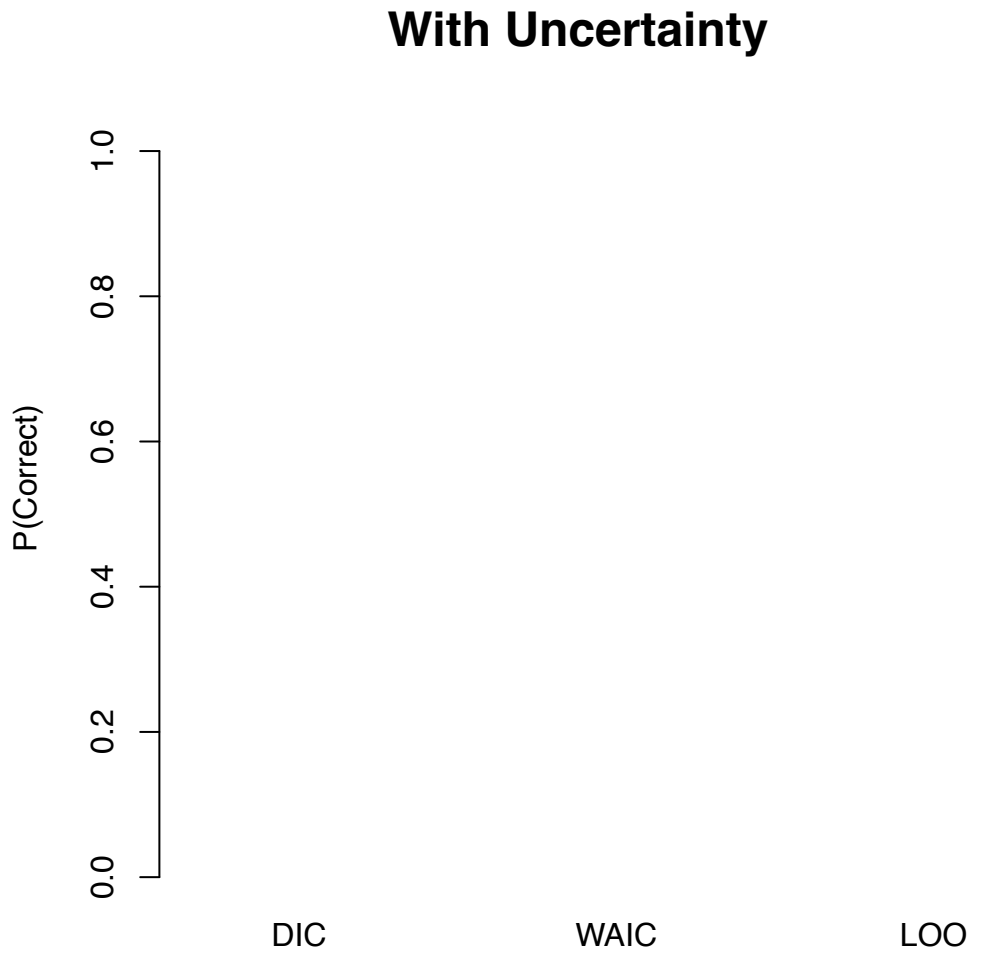
Parameter/Variable	Value
Level 1 Sample Size (i) (e.g., observation, person in a class, trial)	10, 50, 100
Level 2 Sample Size (j) (e.g., classroom, subject, country, cluster)	20, 50, 70
γ_{00}	1
τ_0^2	0.04, 0.16
γ_{10}	0.2, 0.4 *
τ_1^2	0.04, 0.16
γ_{20}	0.2, 0.4
τ_2^2	0
γ_{30}	0
τ_3^2	0
ρ_{01}	0.1
σ^2	Computed per cell to keep total variance at 1

* $\gamma_{20} = 0.2$ explains roughly 4% of the variance in y, as does $\tau_1^2 = 0.04$, because within-cell variance is fixed to 1

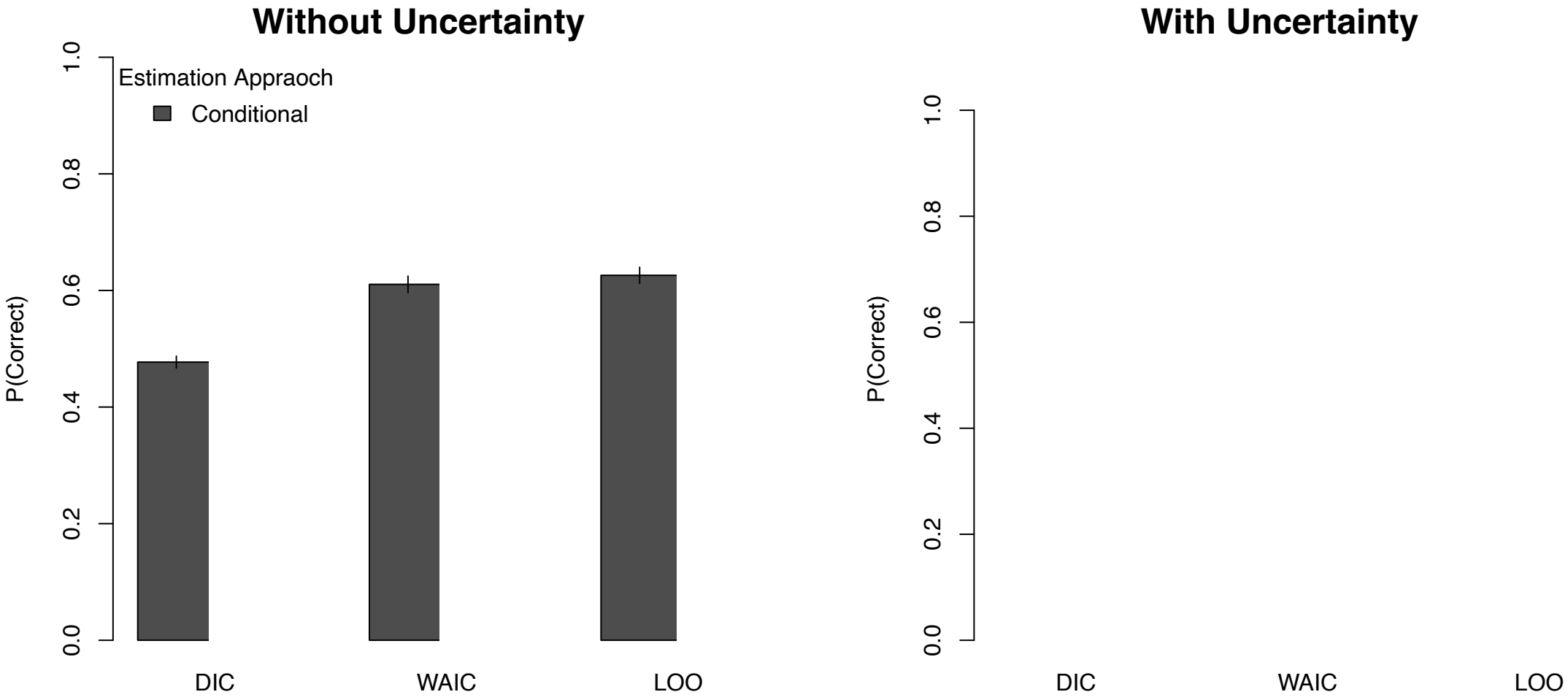
Results

Overall IC selection accuracy

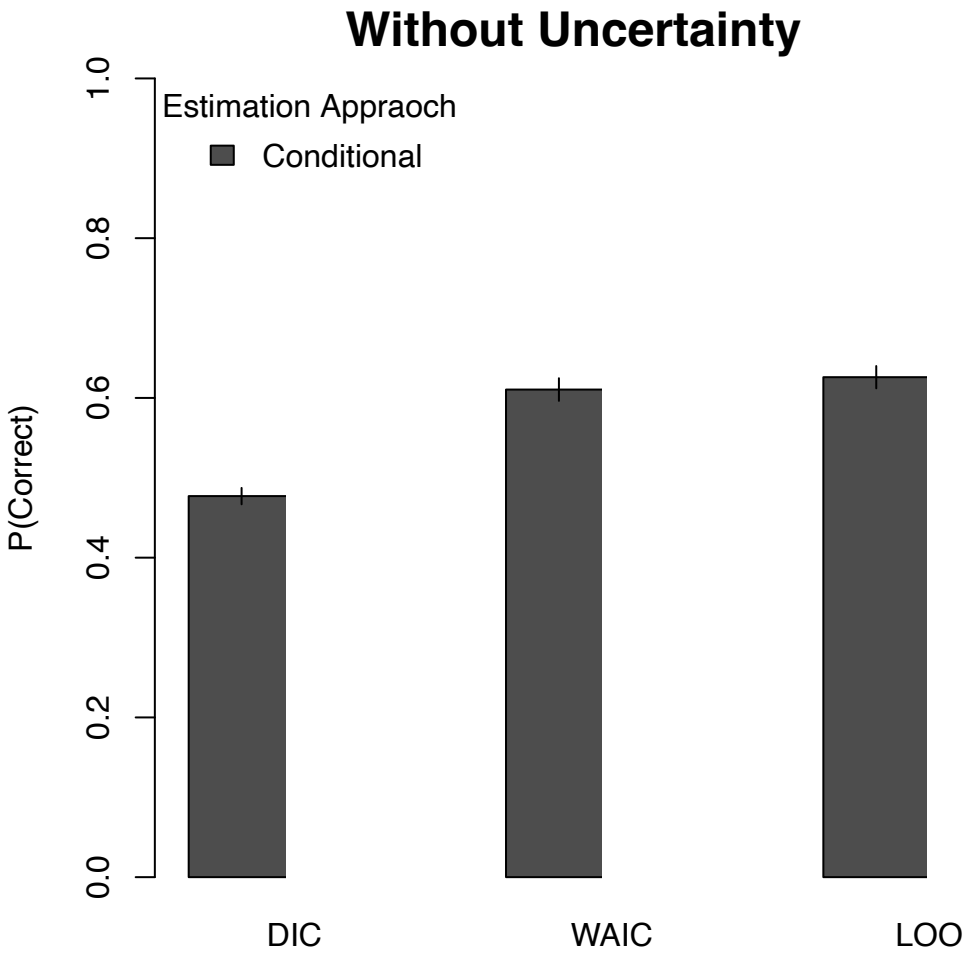


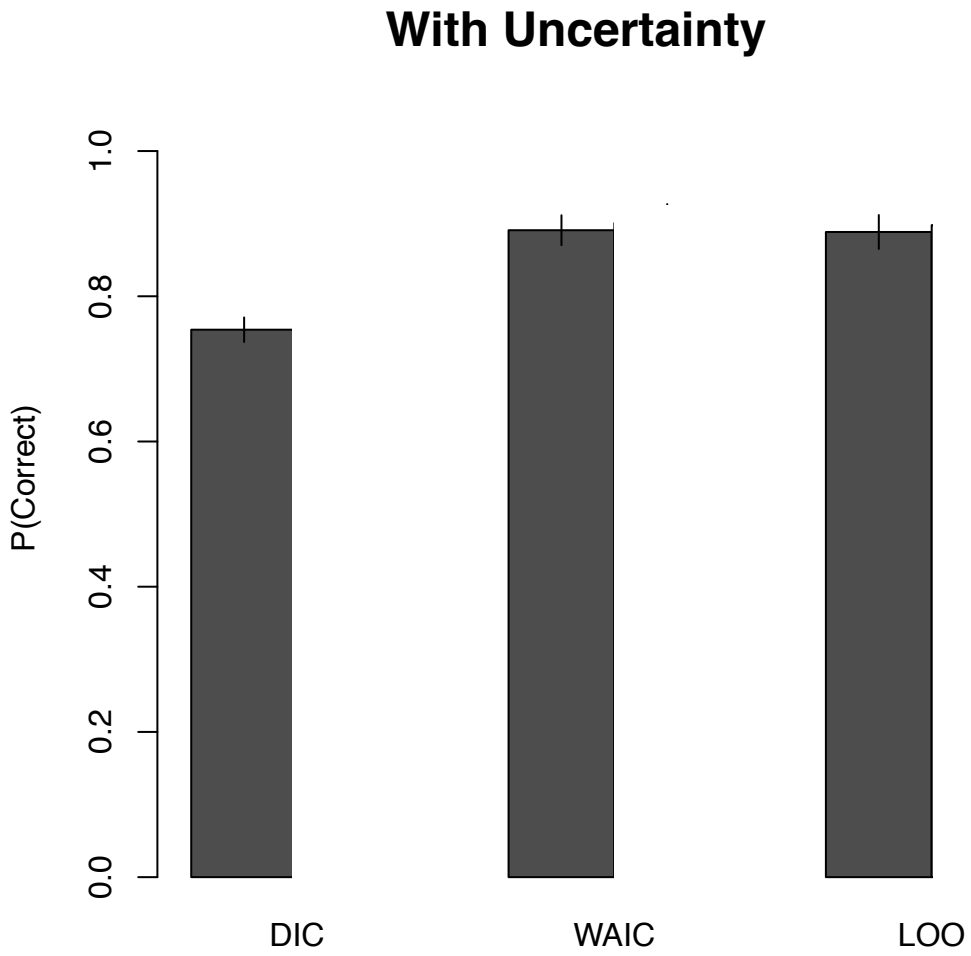


Overall IC selection accuracy

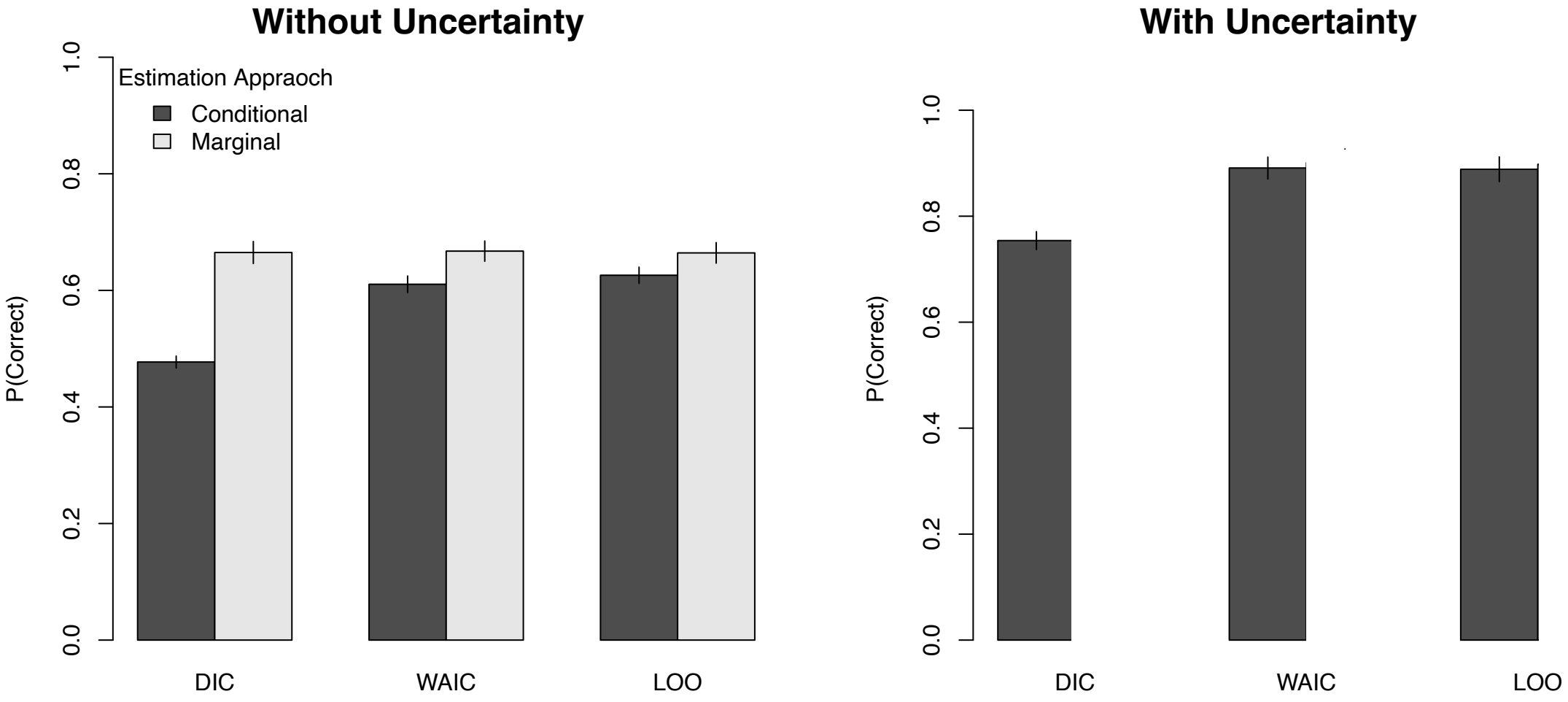


Overall IC selection accuracy

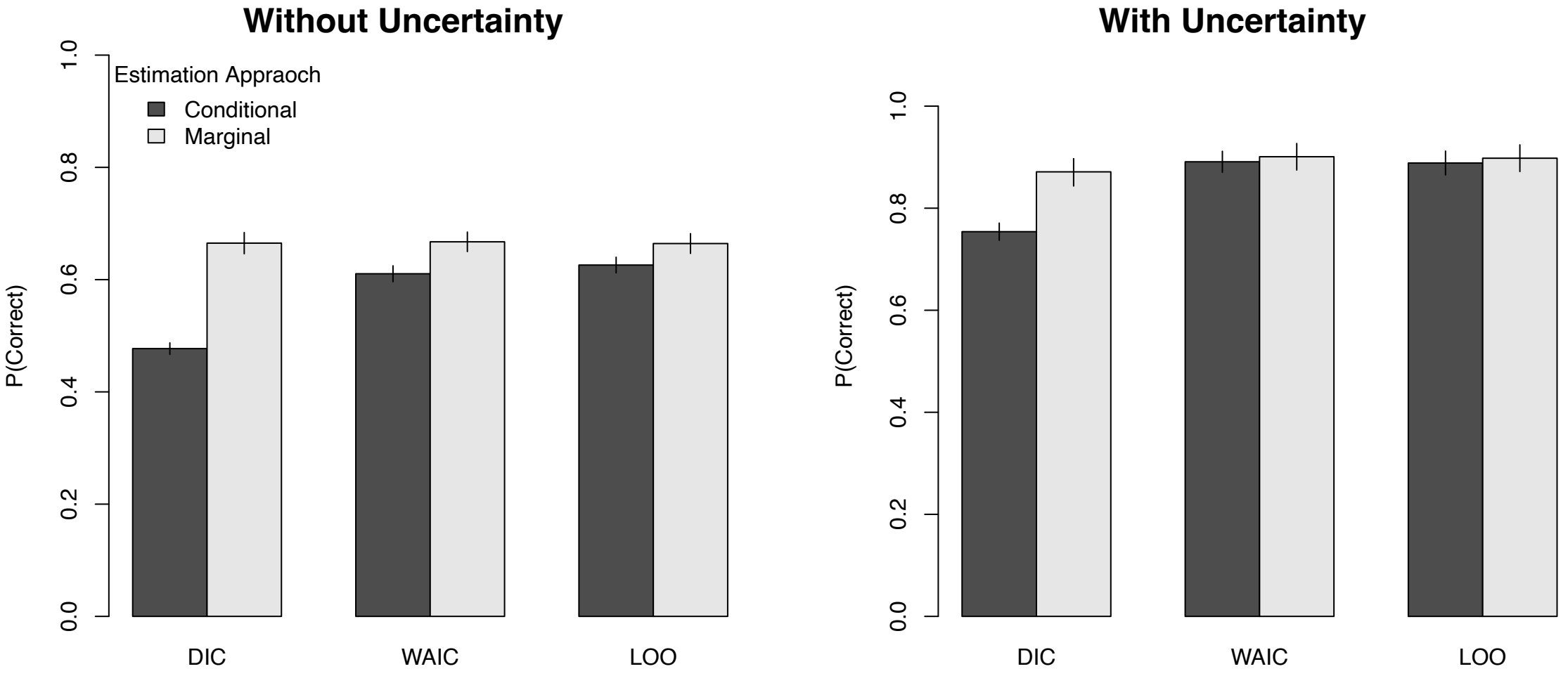




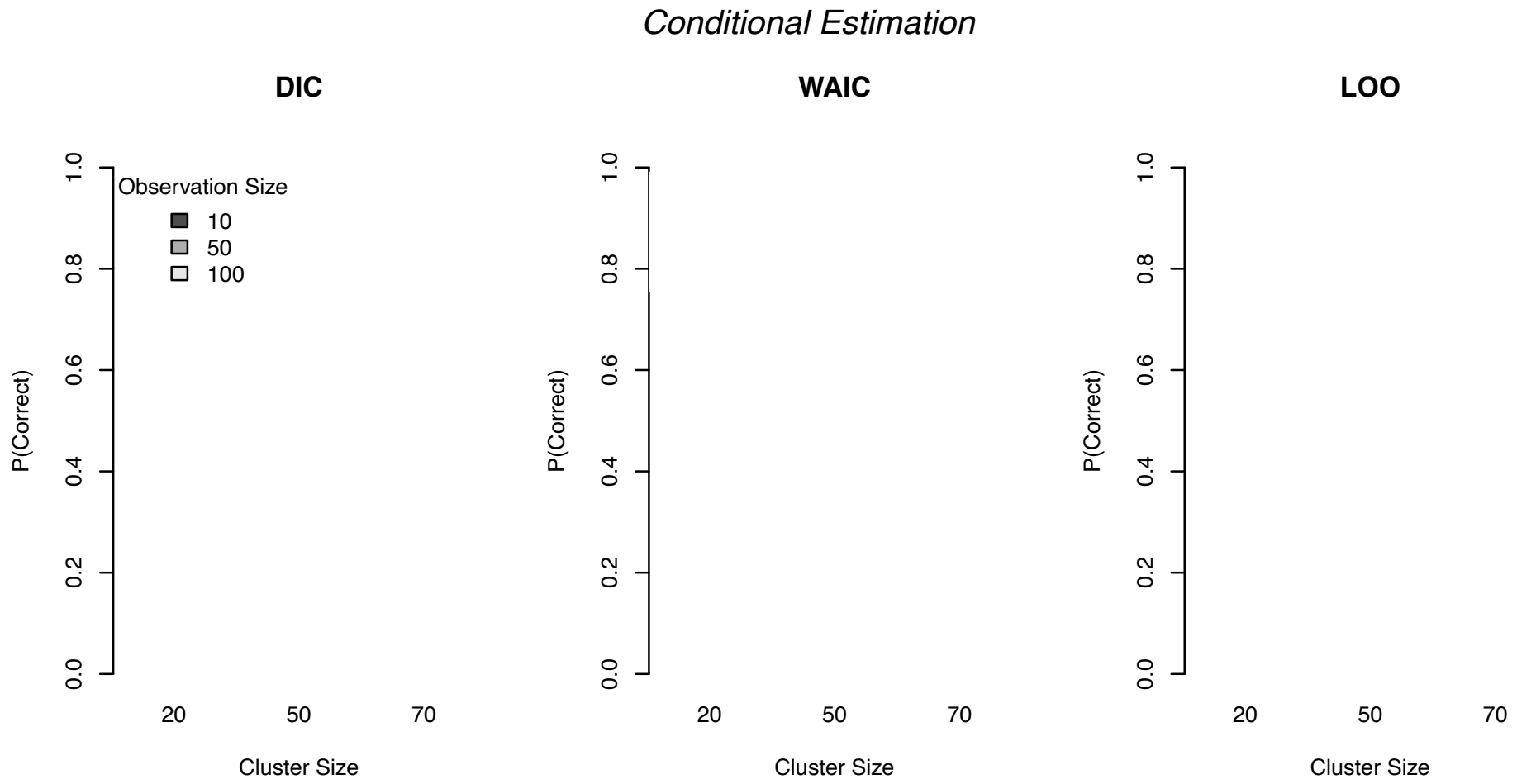
Overall IC selection accuracy



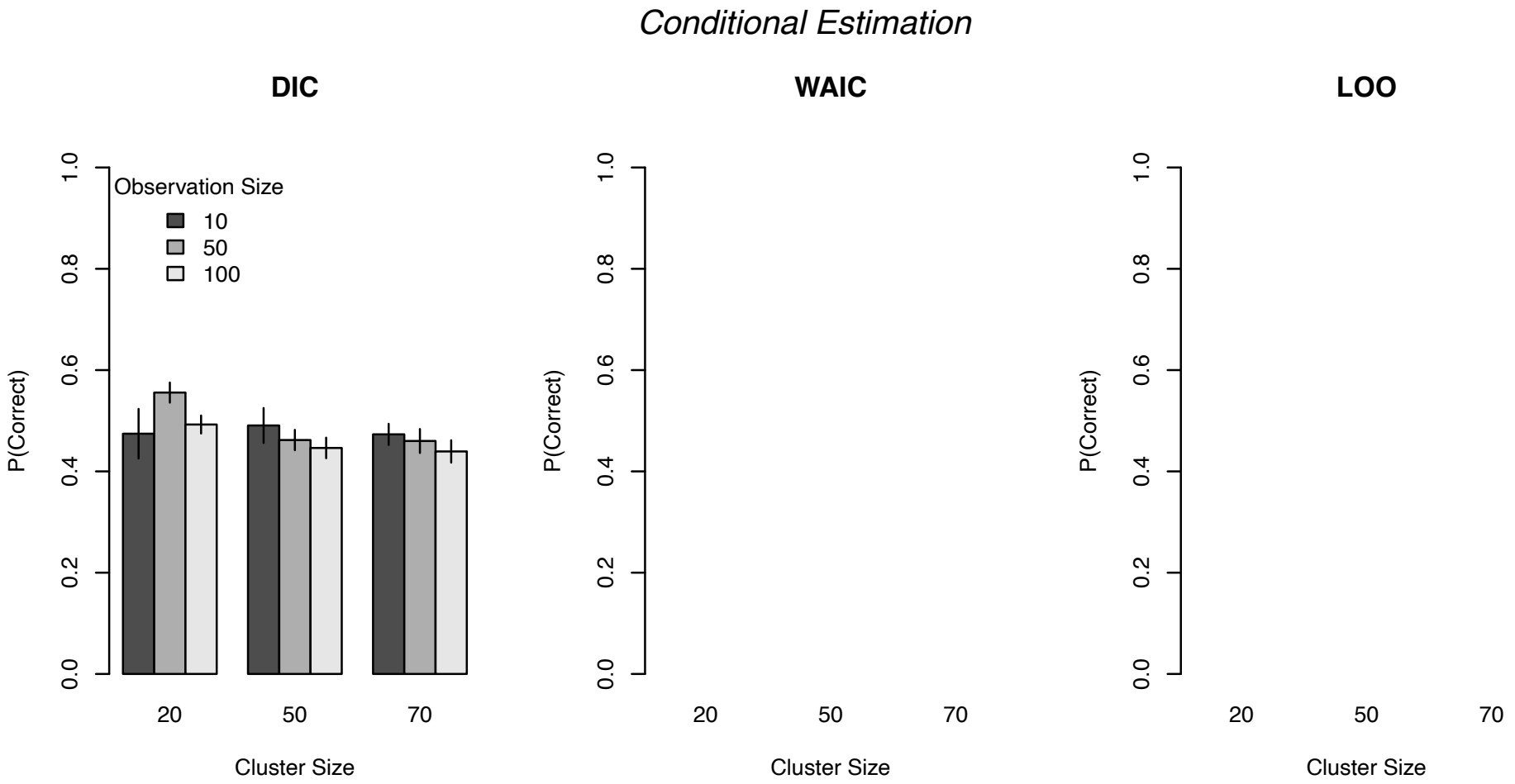
Overall IC selection accuracy



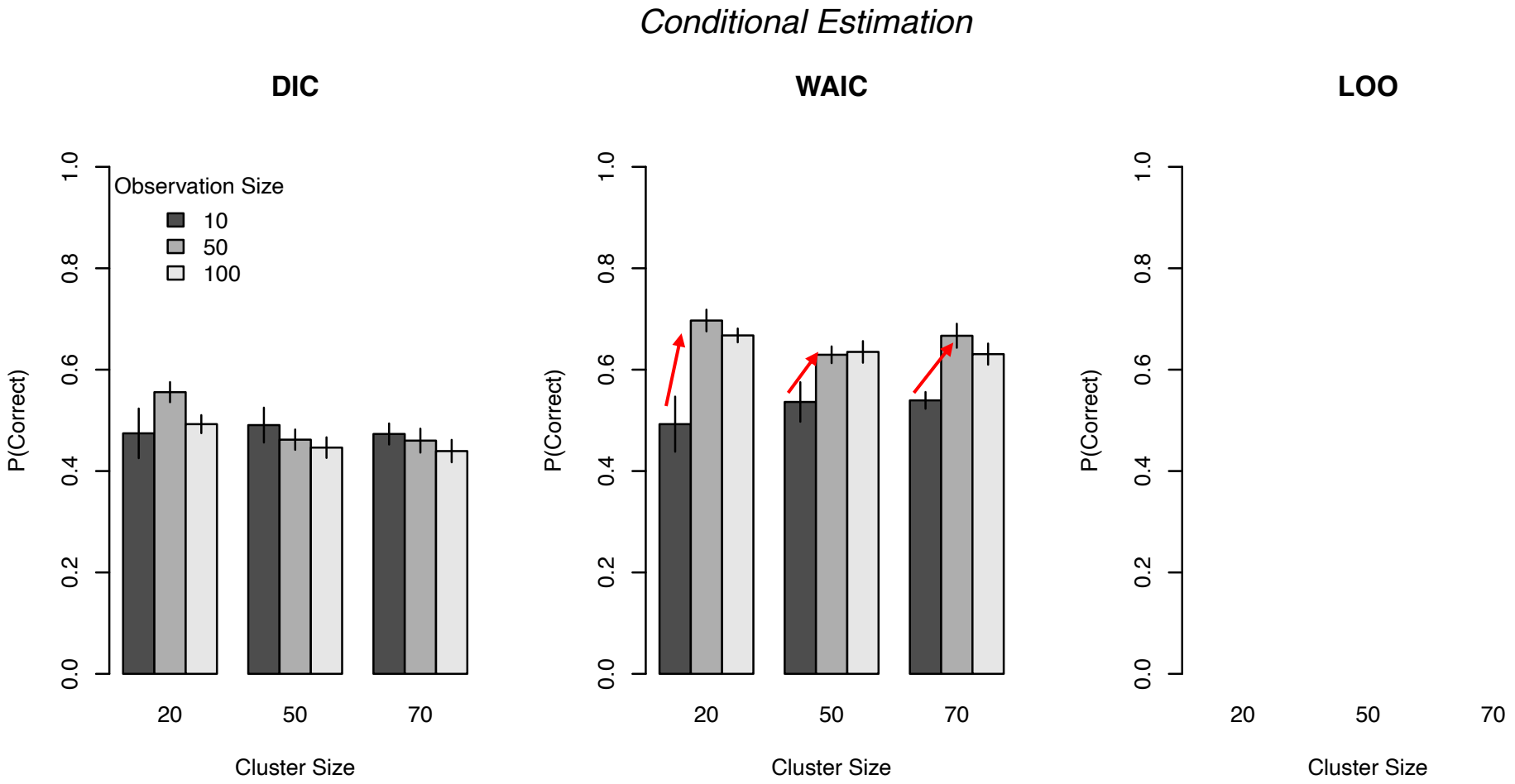
Impact of sample size on accuracy



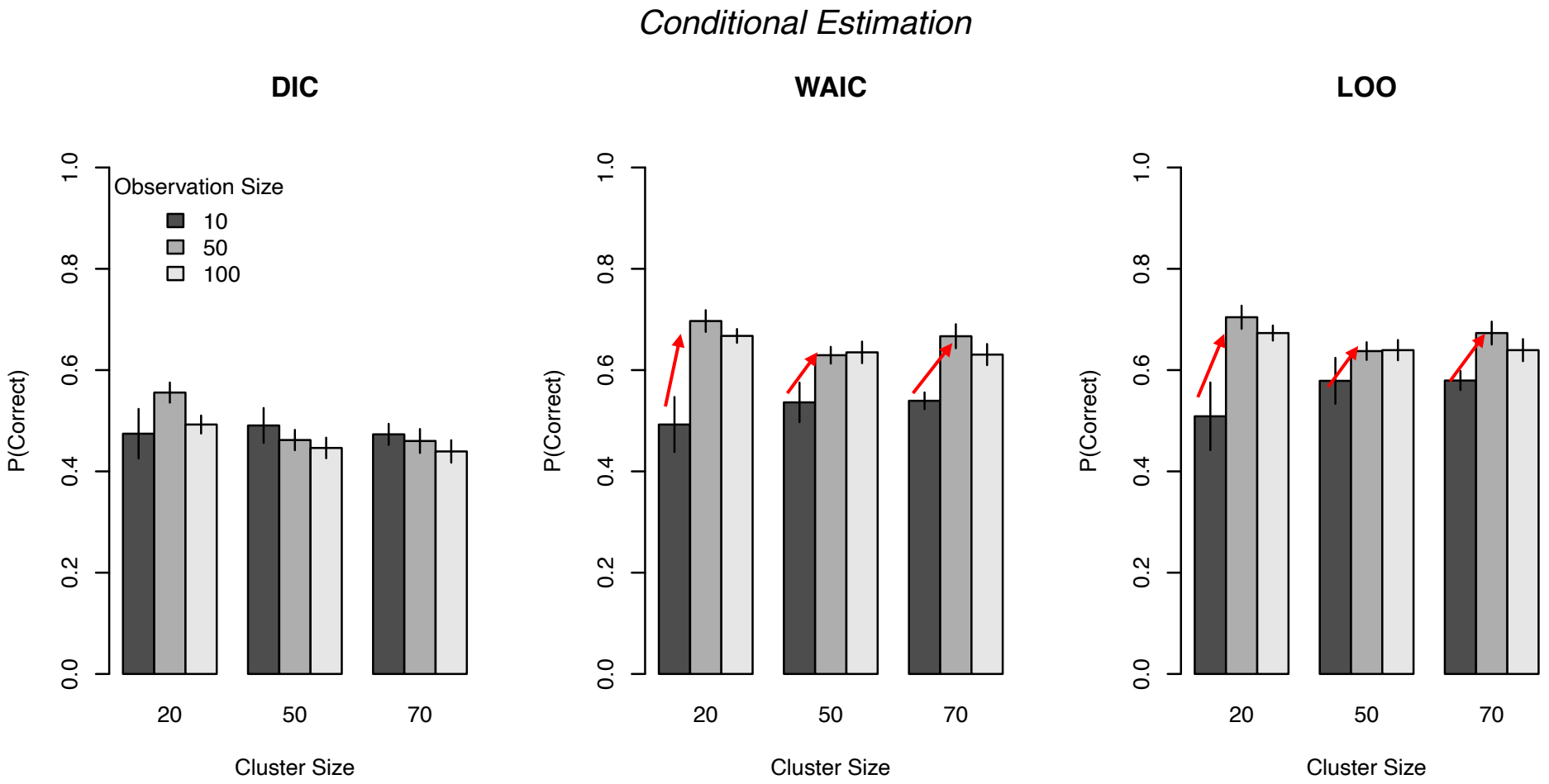
Impact of sample size on accuracy



Impact of sample size on accuracy



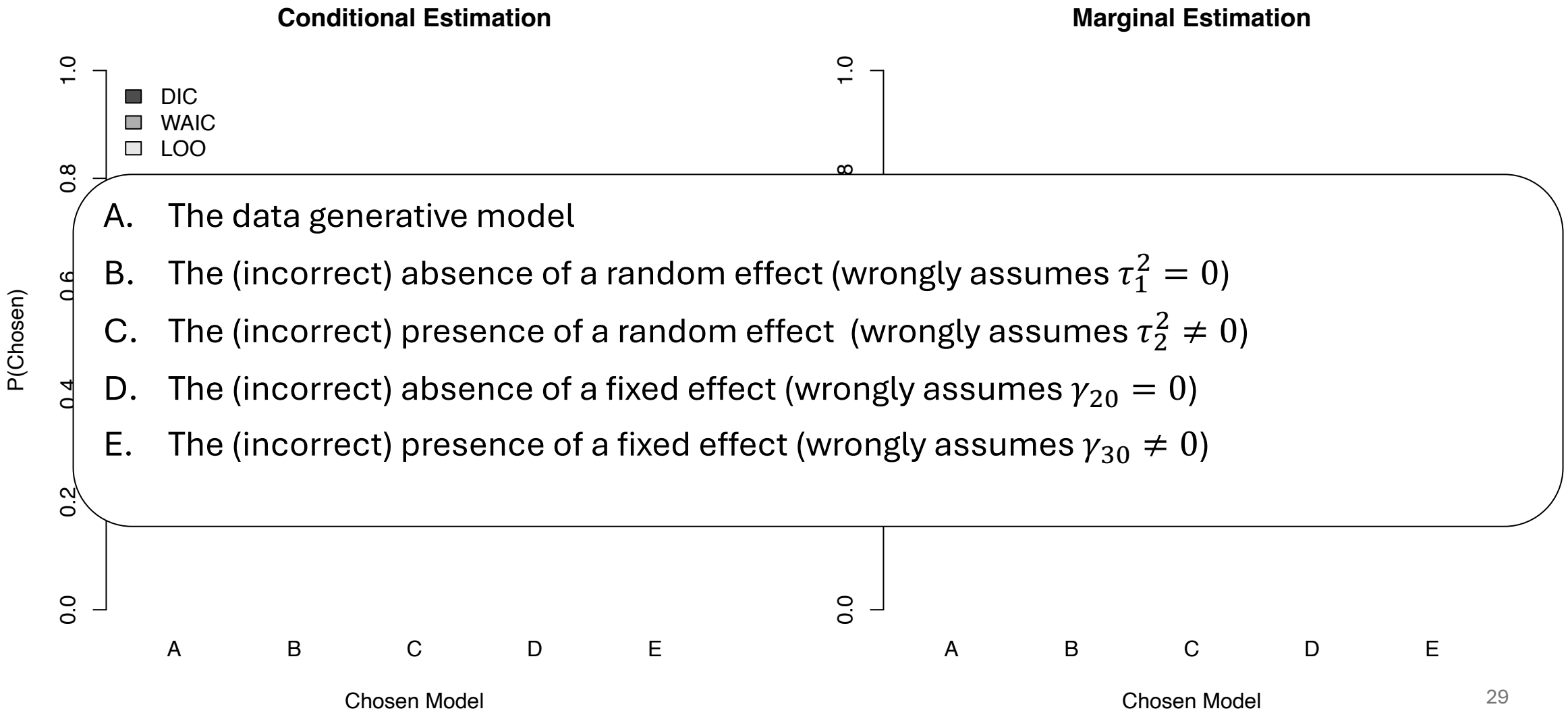
Impact of sample size on accuracy



* DIC shows same pattern as WAIC and LOO under marginal estimation

Other model selection rates

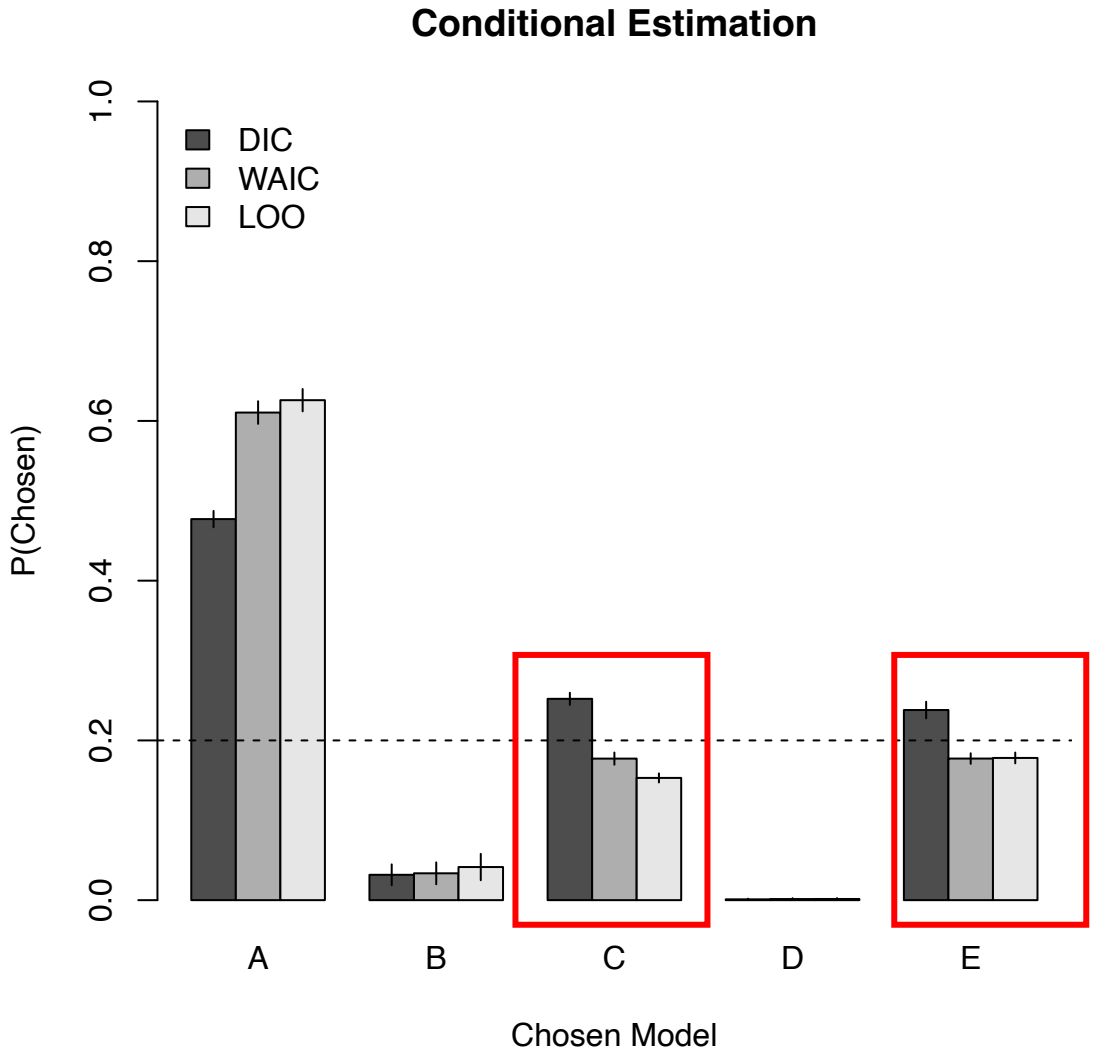
without uncertainty

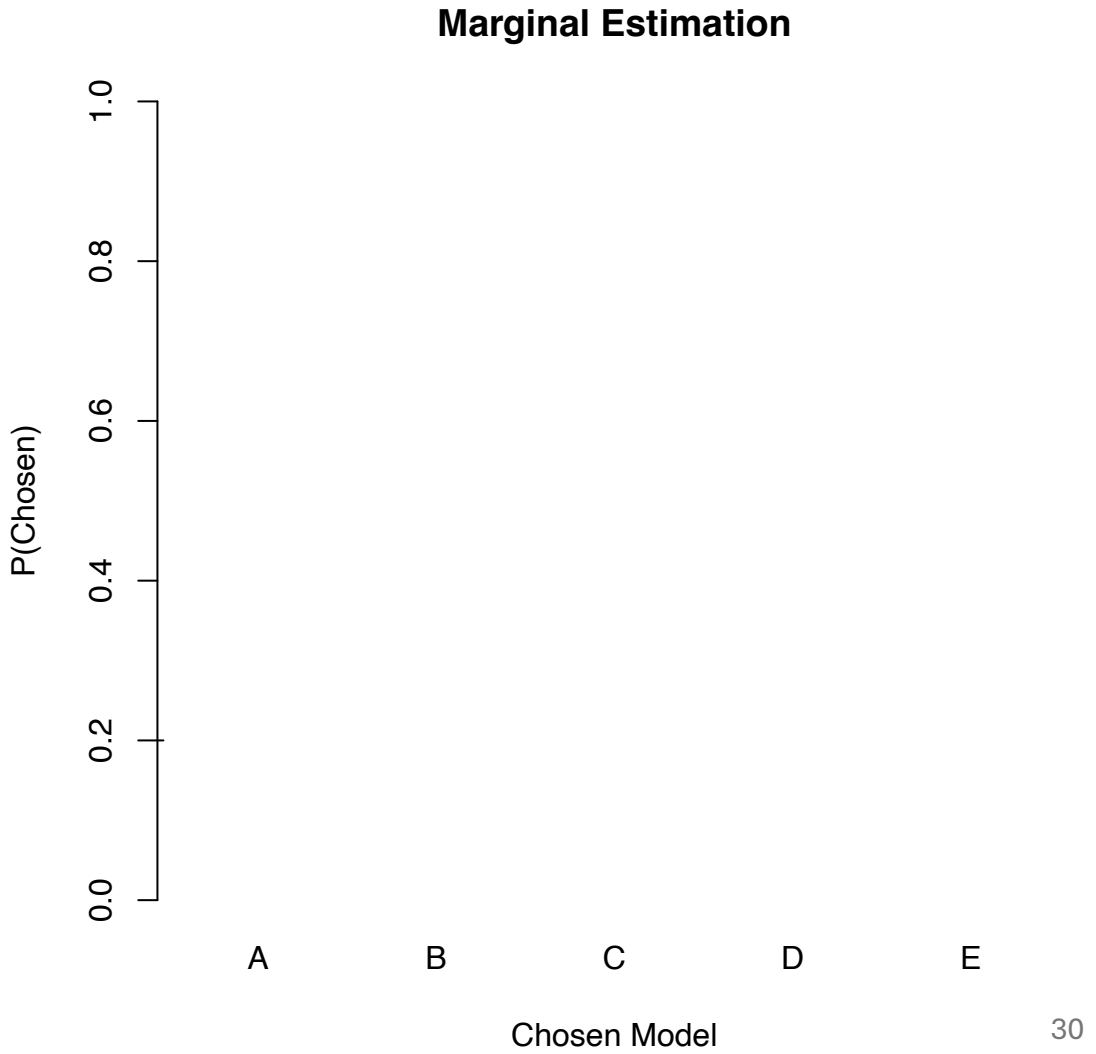


A. The data generative model
B. The (incorrect) absence of a random effect (wrongly assumes $\tau_1^2 = 0$)
C. The (incorrect) presence of a random effect (wrongly assumes $\tau_2^2 \neq 0$)
D. The (incorrect) absence of a fixed effect (wrongly assumes $\gamma_{20} = 0$)
E. The (incorrect) presence of a fixed effect (wrongly assumes $\gamma_{30} \neq 0$)

Other model selection rates

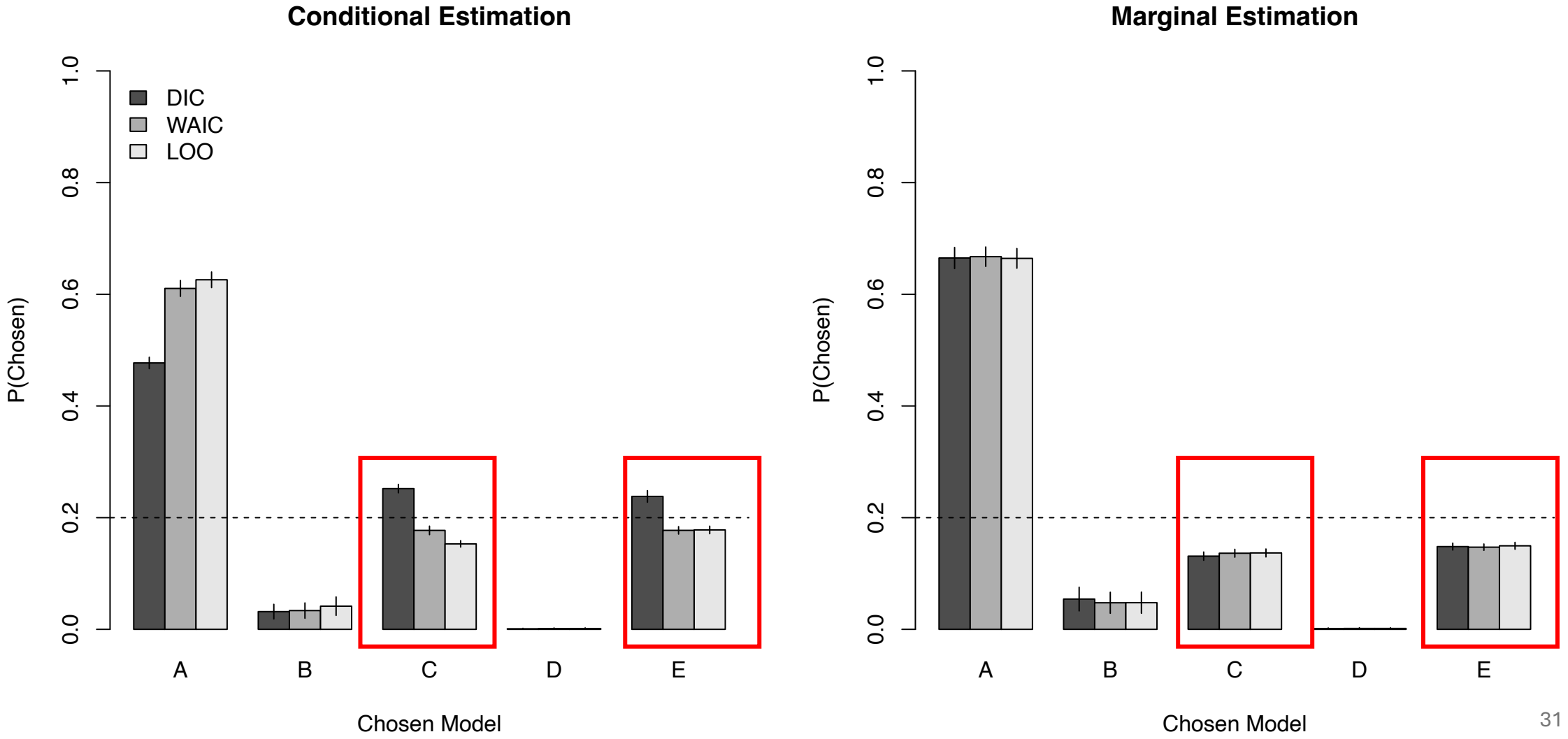
without uncertainty





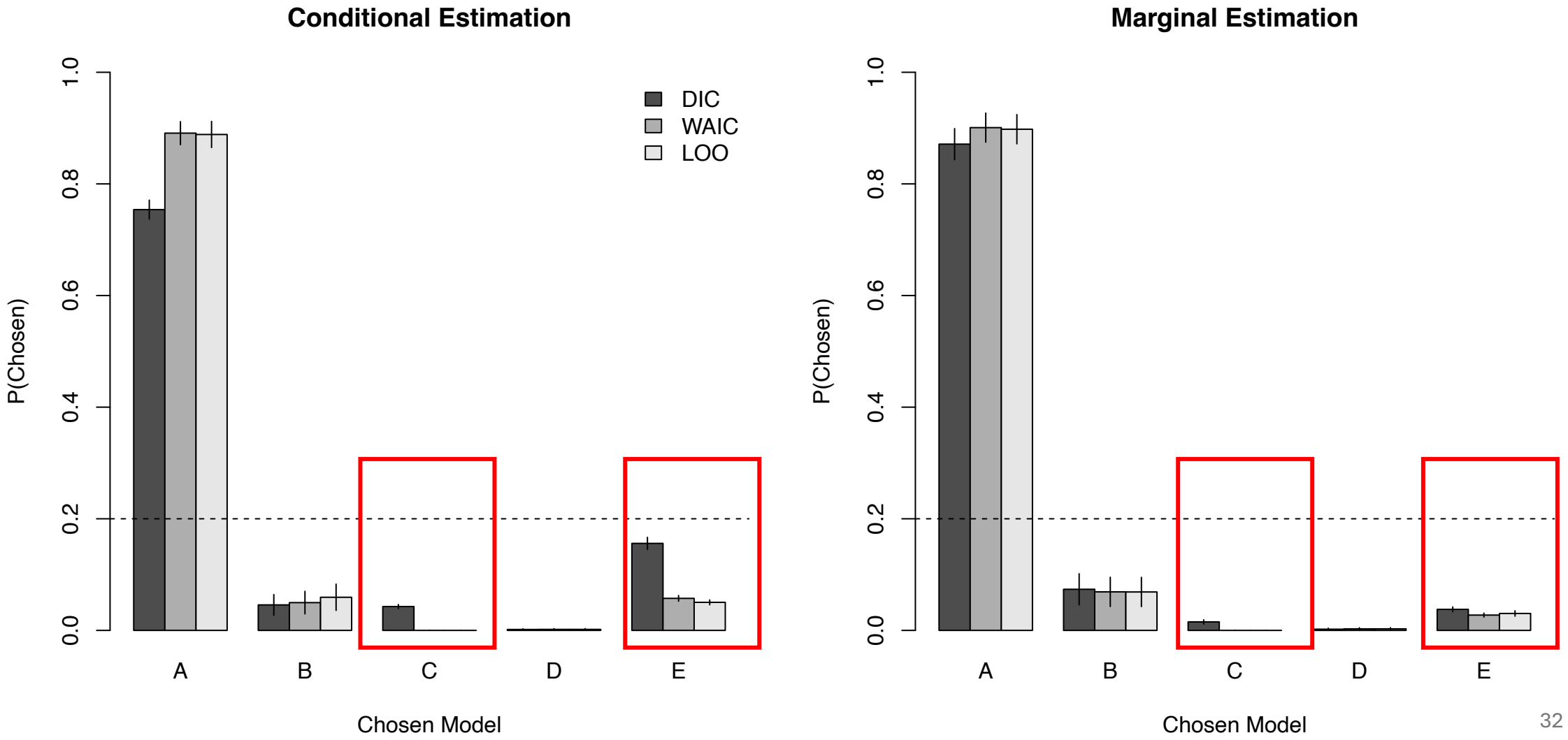
Other model selection rates

without uncertainty



Other model selection rates

with uncertainty



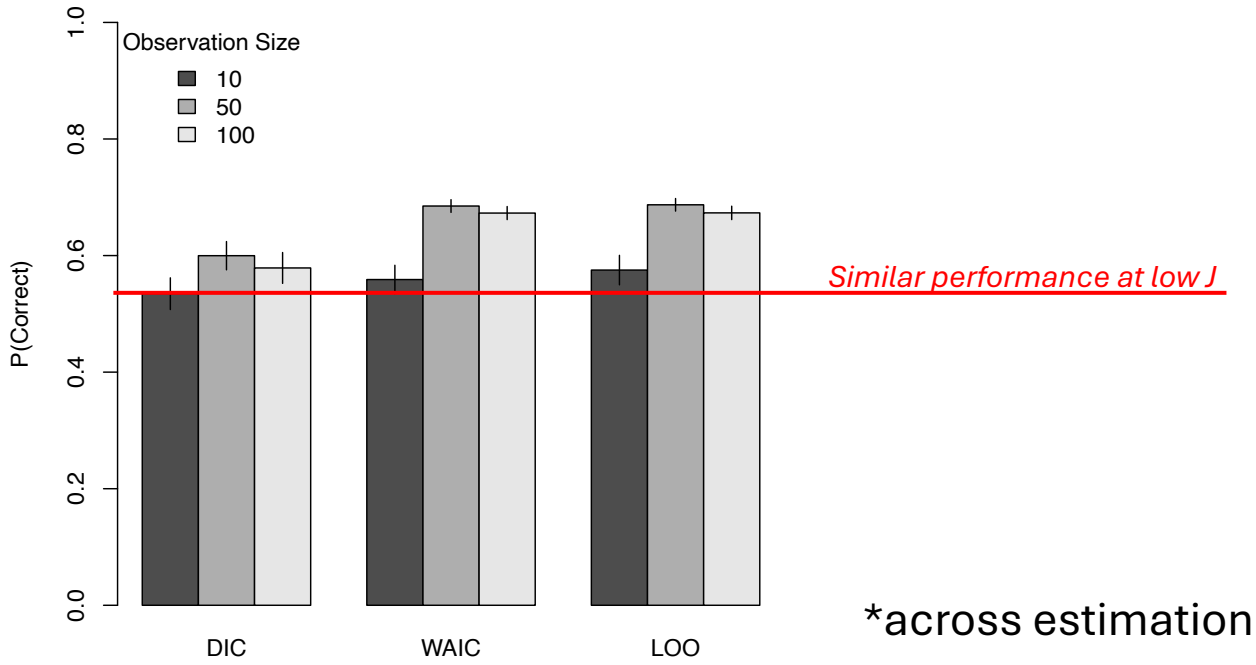
Conclusions

Five main takeaways

1. WAIC and LOO-CV Outperform DIC
   - Consistent with past work in non-multilevel contexts (Ando, 2011)
   - WAIC and LOO-CV outperformed the DIC by more than 10 percentage points in most cases

Five main takeaways

1. WAIC and LOO-CV Outperform DIC
2. Level-1 Sample Size Strongly Affects WAIC and LOO-CV Accuracy
   - While WAIC and LOO-CV outperformed the DIC in the aggregate, when fit to data with a small number of observations (here, 10) per cluster, performance across all three metrics was comparable



*across estimation methods

Five main takeaways

1. WAIC and LOO-CV Outperform DIC
2. Level-1 Sample Size Strongly Affects WAIC and LOO-CV Accuracy
3. The generative model is better recovered under a marginal estimation strategy
   - Using a “lowest value wins” approach, criteria’s accuracy was improved by nearly 10% when computed on marginal (versus conditional) likelihoods.

Five main takeaways

1. WAIC and LOO-CV Outperform DIC
2. Level-1 Sample Size Strongly Affects WAIC and LOO-CV Accuracy
3. The generative model is better recovered under a marginal estimation strategy
4. When Model Selection is Based on a “Lowest Value Wins” Strategy, DIC, WAIC, and LOO-CV May be Prone to Overfitting
   - When ignoring the uncertainty, the three information criteria we considered tended to select models that either overfit the random effects structure (Model C) or the fixed effects structure (Model E).
   - Using this model selection strategy, ***more than a third*** of the decisions may be biased towards overly complex models

Five main takeaways

1. WAIC and LOO-CV Outperform DIC
2. Level-1 Sample Size Strongly Affects WAIC and LOO-CV Accuracy
3. The generative model is better recovered under a marginal estimation strategy
4. When Model Selection is Based on a “Lowest Value Wins” Strategy, DIC, WAIC, and LOO-CV May be Prone to Overfitting
5. Model Selection Uncertainty Should be Considered When Interpreting DIC, WAIC, and LOO-CV
   - Incorporating uncertainty, the overall accuracy on all indices increased, and overfitting rates decreased, by roughly 30 percentage points
   - DIC with uncertainty outperformed WAIC and LOO-CV without uncertainty

Next steps

- A subsequent study in which
  - Predictors are correlated
  - Sample sizes are smaller
  - ICCs are larger

Carl Falk, McGill University

Read our preprint

Ken A. Fujimoto, Loyola University Chicago

Thank you for listening ☺

(I am on the job market)

McGill UNIVERSITY

NSERC CRSNG

Calcul Québec

Simulation values

- Fit with RStan (Stan Development Team, 2024)
  - *brms* for conditional estimation
  - Custom stan code for marginal estimation
- 2 chains per model
- 3000 samples/chain (1000 burn-in)
- 288 cells
- 100 samples per cell

IC Calculation details

DIC

$$\text{DIC} = -2\ln p\left(\boldsymbol{y}|\bar{\bar{\boldsymbol{\theta}}}\right) + 2p_D \qquad \text{Eq. 1}$$

where $\ln p\left(\boldsymbol{y}|\bar{\bar{\boldsymbol{\theta}}}\right)$ is the likelihood of the data based on the posterior mean of the parameters (i.e., $\bar{\bar{\boldsymbol{\theta}}}$) and p_d is a data-driven bias correction term that is calculated through:

$$p_D = -2\left(\frac{1}{S}\sum_{s=1}^{S}\ln p(\boldsymbol{y}|\boldsymbol{\theta}^s) - \ln p\left(\boldsymbol{y}\middle|\bar{\bar{\boldsymbol{\theta}}}\right)\right) \qquad \text{Eq. 2}$$

where the first part inside the parentheses is the posterior mean of the deviances calculated using the sth sampled values for the parameters (i.e., $\boldsymbol{\theta}^s$) during the MCMC sampling process, with $s = 1, \ldots S$ and S being the total saved number of sampled values from the posterior distribution. In other words, p_D is the "mean deviance minus the deviance at the posterior means of the parameters"—that is, it captures the degree to which the likelihood of the data under $\bar{\bar{\boldsymbol{\theta}}}$ deviates from the average likelihood of the data under all posterior draws. If this value is large, there is a large discrepancy between the average likelihood and the likelihood of the data under the average posterior parameter estimates, suggesting the model is overfitting the data. Conversely, if p_D is small, it suggests that the posterior is precisely centred around $\bar{\bar{\boldsymbol{\theta}}}$ and thus the number of effective parameters is small. Apart from criticisms about its efficacy as a good metric for model selection (Ando, 2011), the DIC has also been criticized as not being "fully Bayesian" because its estimate of the likelihood relies on point estimates, $\bar{\bar{\boldsymbol{\theta}}}$, rather than using all the information contained in the posterior, which is to many Bayesians a key advantage of the

WAIC

Watanabe-Akaike information criterion (WAIC)

The WAIC is a "fully-Bayesian" approach to estimating the out-of-sample predictive accuracy because it is based on the full posterior predictive density of each point, which when taking the log of it, we have the log pointwise posterior predictive density (lppd), compared with the DIC, which depends on a point estimate of the predictive likelihood. By log pointwise posterior predictive density, we mean that the log probability of each data point y_{ij} given each $\boldsymbol{\theta}^s$ and is obtained through:

$$lppd = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \ln \frac{1}{S}\sum_{s=1}^{S} p(y_{ij}|\boldsymbol{\theta}^s) \qquad \text{Eq. 3}$$

Because the WAIC depends on the lppd, the likelihood of each data point is evaluated using information from the entire posterior distribution, rather than only its central tendency. Similarly, the penalty term of the WAIC, p_W, is applied pointwise to each data point, y_{ij}, as[2]:

$$p_W = \sum_{j=1}^{J}\sum_{i=1}^{n_j} (V_{s=1}^{S} \ln p(y_{ij} \mid \boldsymbol{\theta}^s) \qquad \text{Eq. 4}$$

where, V represents the sample variance computed over the posterior, $V_{s=1}^{S} \ln p(y_{ij} \mid \boldsymbol{\theta}^s) = \frac{1}{S-1}\sum_{s=1}^{S}\left(\ln p(y_{ij}|\boldsymbol{\theta}^s) - \overline{\ln p(y_{ij} \mid \boldsymbol{\theta}^s)}\right)^2$. These terms are then combined and multiplied by -2 in order to put the WAIC on a deviance scale, i.e., on the same scale as the DIC:

$$\text{WAIC} = -2(lppd - p_W) \qquad \text{Eq. 5}$$

LOO

Leave-One-Out Information Criterion (LOOIC)

As mentioned in the introduction, the goal of information criteria is to estimate the predictive accuracy of a model in a new sample—that is, when applied to new data, what is the expected likelihood of the new data under this model? A simple way to accomplish this would be to split the data into two sets, a training and a testing set, to estimate the model on the training set, and to predict the data in the testing set, noting the discrepancy directly.

Leave-one-out cross-validation (LOO-CV) implements a version of this concept. For a given data point, y_{ij}, LOO-CV trains the model on all other data points in the sample, $\boldsymbol{y}_{-ij}$, and the predictive likelihood will be computed on y_{ij}:

$$\text{LOO-CV} = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \ln \frac{1}{S}\sum_{s=1}^{S} p(y_{ij}|\boldsymbol{y}_{-ij}, \boldsymbol{\theta}^s) \qquad \text{Eq. 6}$$

where $p(y_{ij}|\boldsymbol{y}_{-ij}, \boldsymbol{\theta}^s)$ is the predictive density for y_{ij} given the sth sampled values from the posterior distribution based on only $\boldsymbol{y}_{-ij}$. However, computing $p(y_{ij}|\boldsymbol{y}_{-ij}, \boldsymbol{\theta}^s)$ in this way is often computationally prohibitive for data with many observations when Bayesian estimation is used. It is difficult to efficiently sample from the posterior with a single data point removed, $p(\boldsymbol{\theta}^s|\boldsymbol{y}_{-ij})$, and repeat the process for all data points. To circumvent this issue, Vehtari et al. (2017) proposed a method to approximate the LOO-CV from the observed posterior draws, using importance sampling. In short, importance sampling helps researchers learn about features of some target distribution by drawing samples from a second distribution that may be easier to sample from. Estimates of the target distribution are weighted by the ratio of the likelihoods of the distributions, r. Here, this technique is leveraged to infer the LOO posterior, $p(\boldsymbol{\theta}^s|\boldsymbol{y}_{-ij})$, from the full data posterior $p(\boldsymbol{\theta}^s|\boldsymbol{y})$, using raw weights, $r_{ijs} = \frac{1}{p(y_{ij}|\boldsymbol{\theta}^s)}$, which can be used to evaluate the log predictive density at the left-out data point, as:

$$p(y_{ij}|\boldsymbol{y}_{-ij}) \approx \frac{\sum_{s=1}^{S} r_{ij}^s \, p(y_{ij}|\boldsymbol{\theta}^s)}{\sum_{s=1}^{S} r_{ij}^s} \qquad \text{Eq. 7}$$

However, the values produced by Eq. 7 can be unreliable in some circumstances because the r_{ijs} can be unstable. This can occur when the variance and shape of the full data posterior, $p(\boldsymbol{\theta}^s|D)$, differs appreciably from that of the LOO posterior, $p(\boldsymbol{\theta}^s|D_{-i})$. For example, if $p(\boldsymbol{\theta}^s|\boldsymbol{y})$ has a much smaller variance than $p(\boldsymbol{\theta}^s|\boldsymbol{y}_{-ij})$, r_{ij}^s values will be very large when sampling the tails of $p(\boldsymbol{\theta}^s|\boldsymbol{y})$, thus mischaracterizing (i.e., shrinking) the true spread of values in $p(\boldsymbol{\theta}^s|\boldsymbol{y}_{-ij})$. To overcome this issue, Vehtari et al. (2022) applied a smoothing procedure to the extreme raw ratios (r_{ij}^s), with this procedure based on the Pareto distribution, leading to an updated vector of weights, w_{ij}^s. Details are presented in Vehtari et al. (2017), but in short these weights counteract the potential bias from using raw weights. Once the modified weights are obtained, the LOO expected log pointwise predictive density can be computed as:

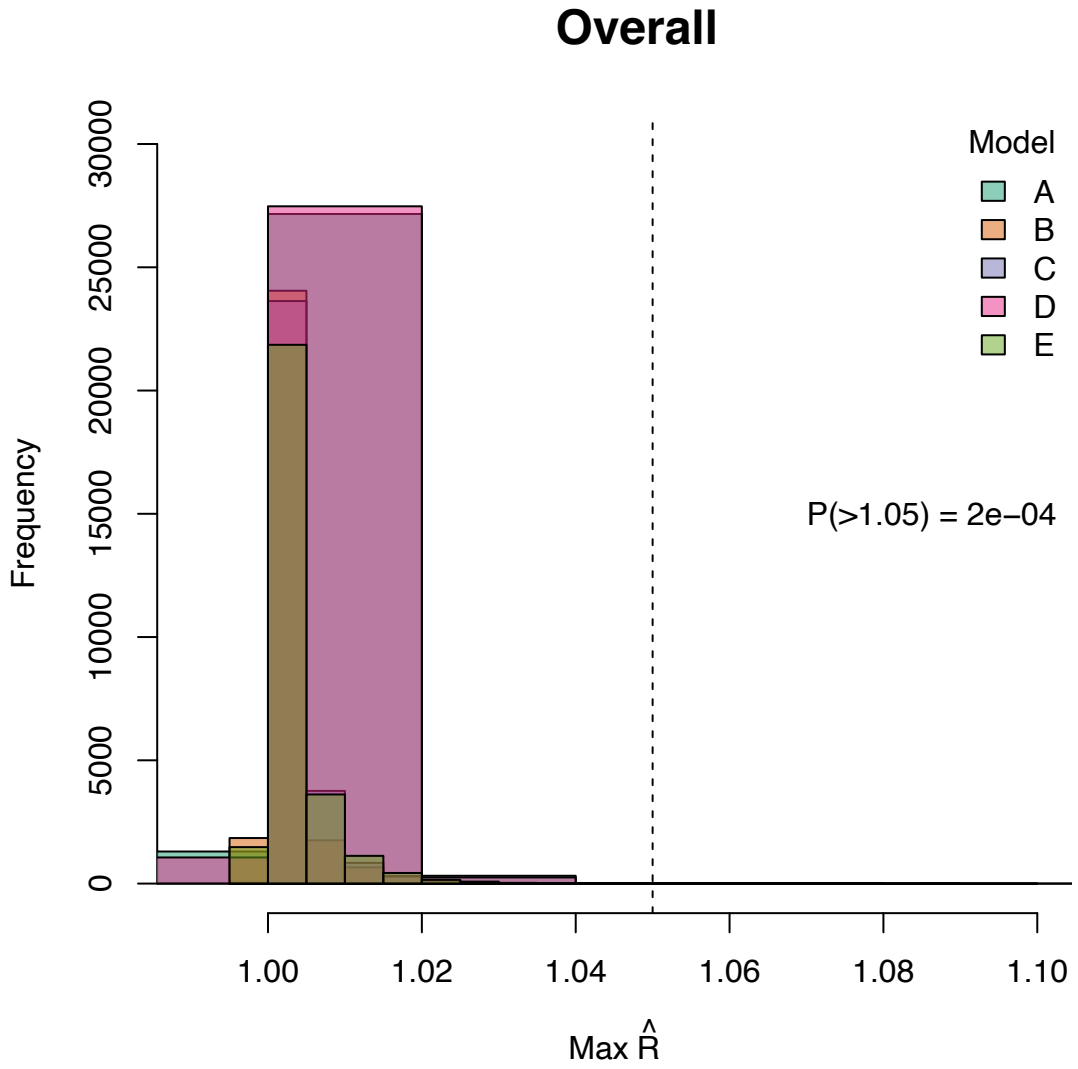
$$elppd = \sum_{j=1}^{J}\sum_{i=1}^{n_j} \ln\left(\frac{\sum_{s=1}^{S} w_{ij}^s p(y_{ij}|\boldsymbol{\theta}^s)}{\sum_{s=1}^{S} w_{ij}^s}\right) \qquad \text{Eq. 8}$$

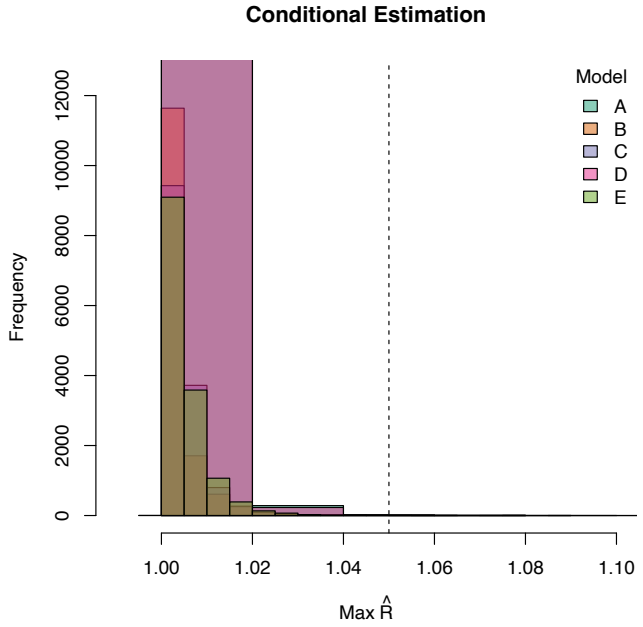
which when multiplied by -2 yields the LOO-CV on the deviance scale:

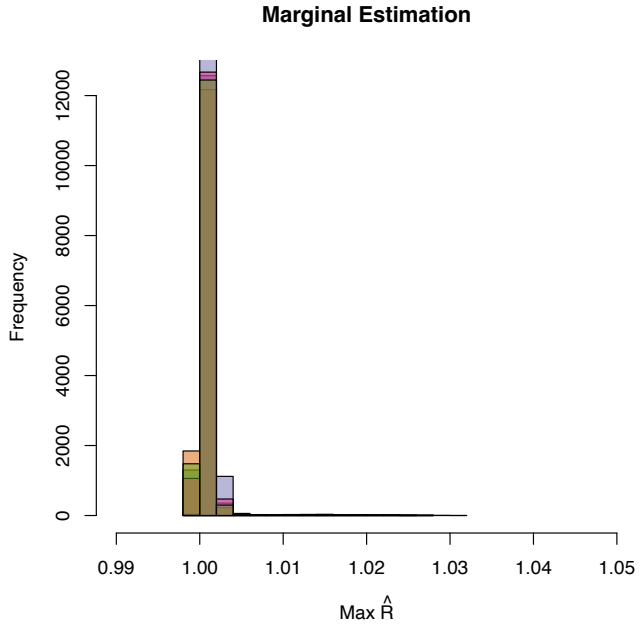
$$\text{LOO-CV} = -2elppd \qquad \text{Eq. 9}$$

Model diagnostics

Model diagnostics (rhat)

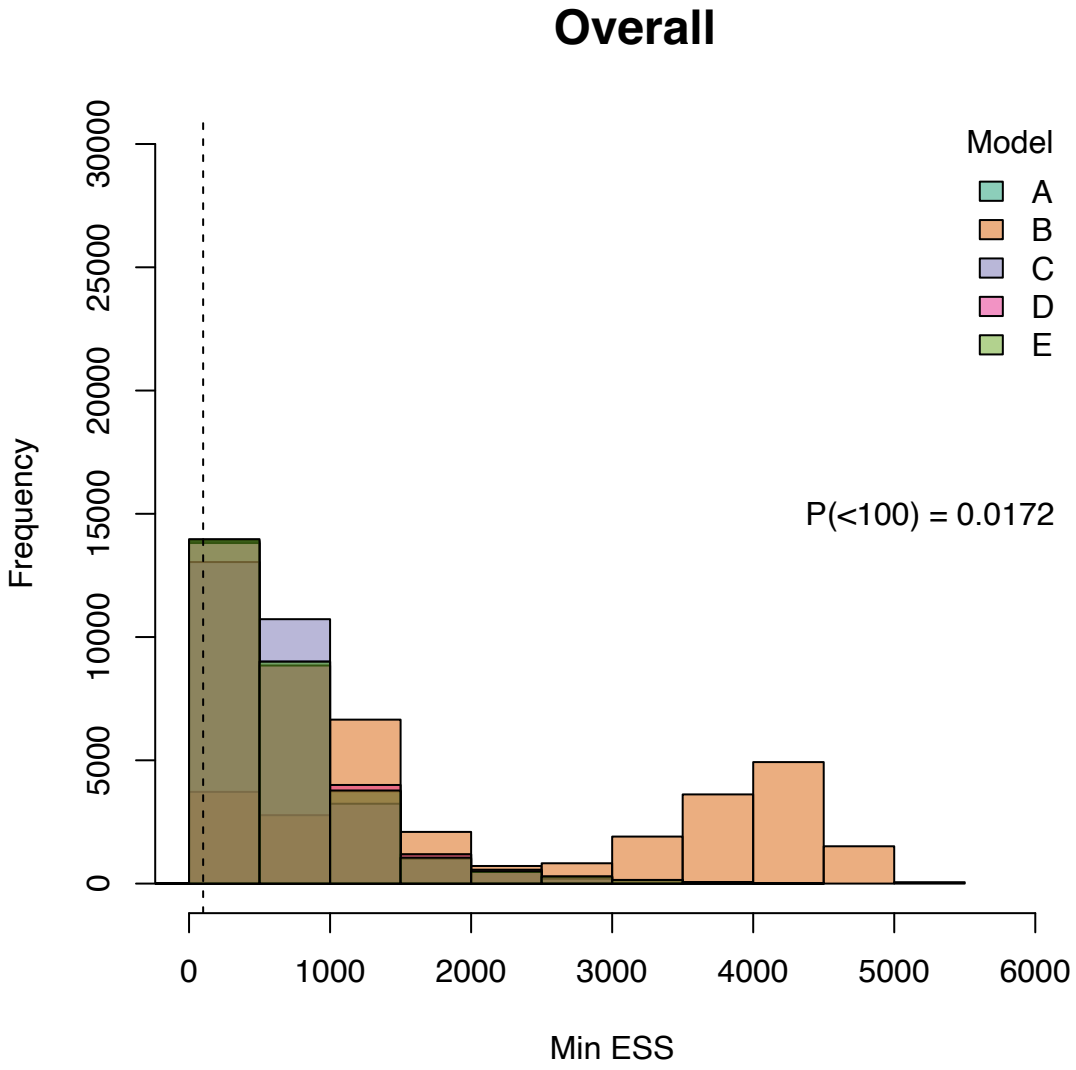


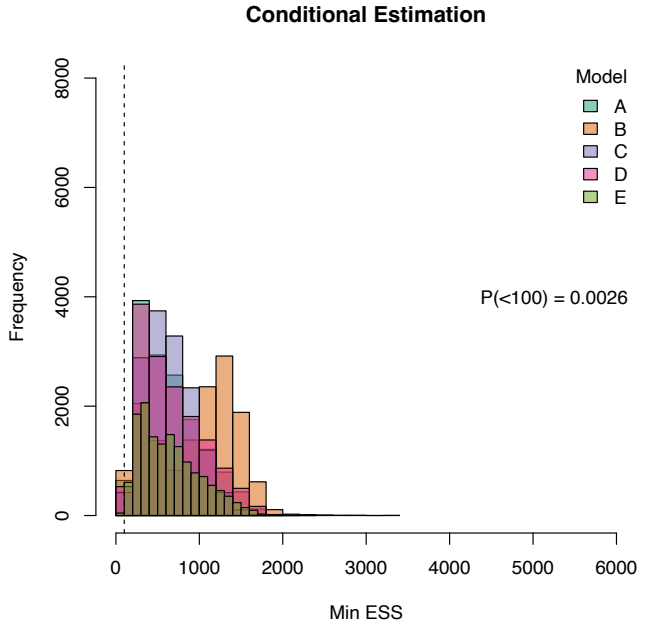


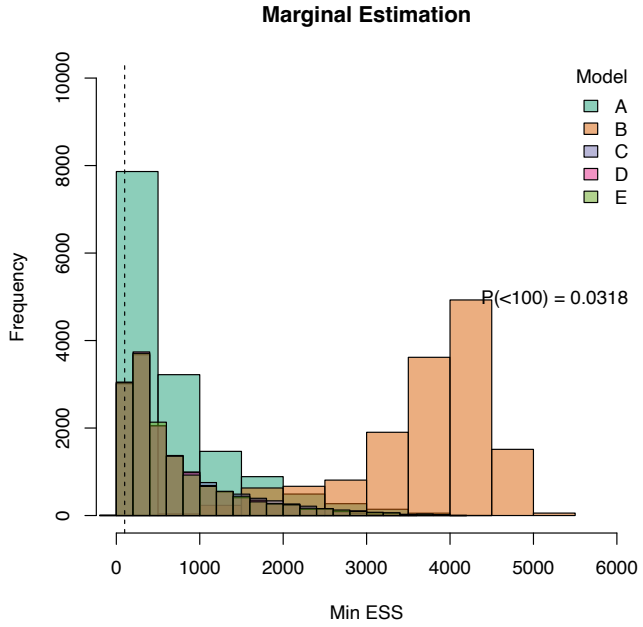


Good convergence for the vast majority of models, especially for marginal estimation (likely owing to fewer random effects to estimate)

Model diagnostics (ESS)



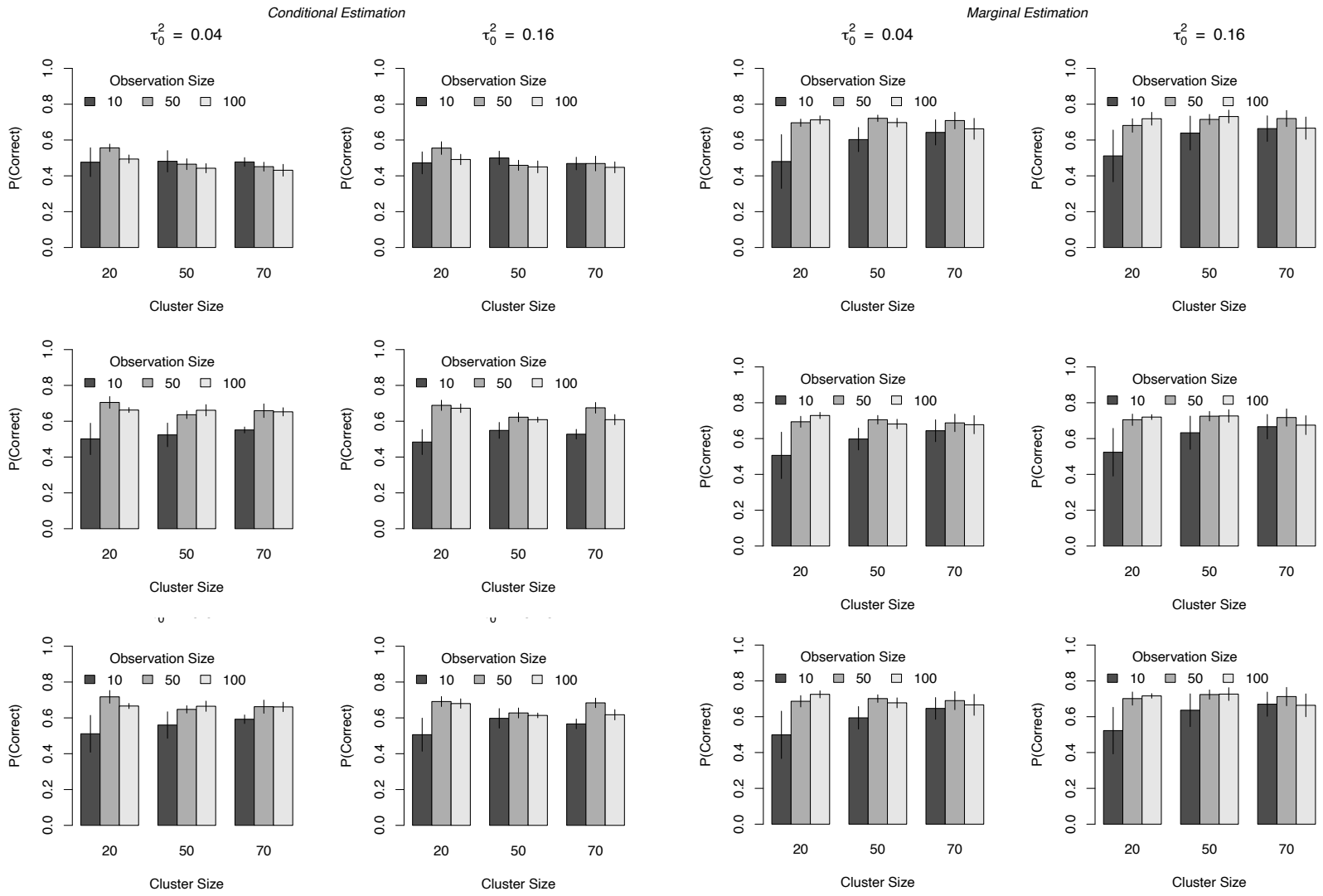




Good convergence for the vast majority of models, especially for marginal estimation (likely owing to fewer random effects to estimate)

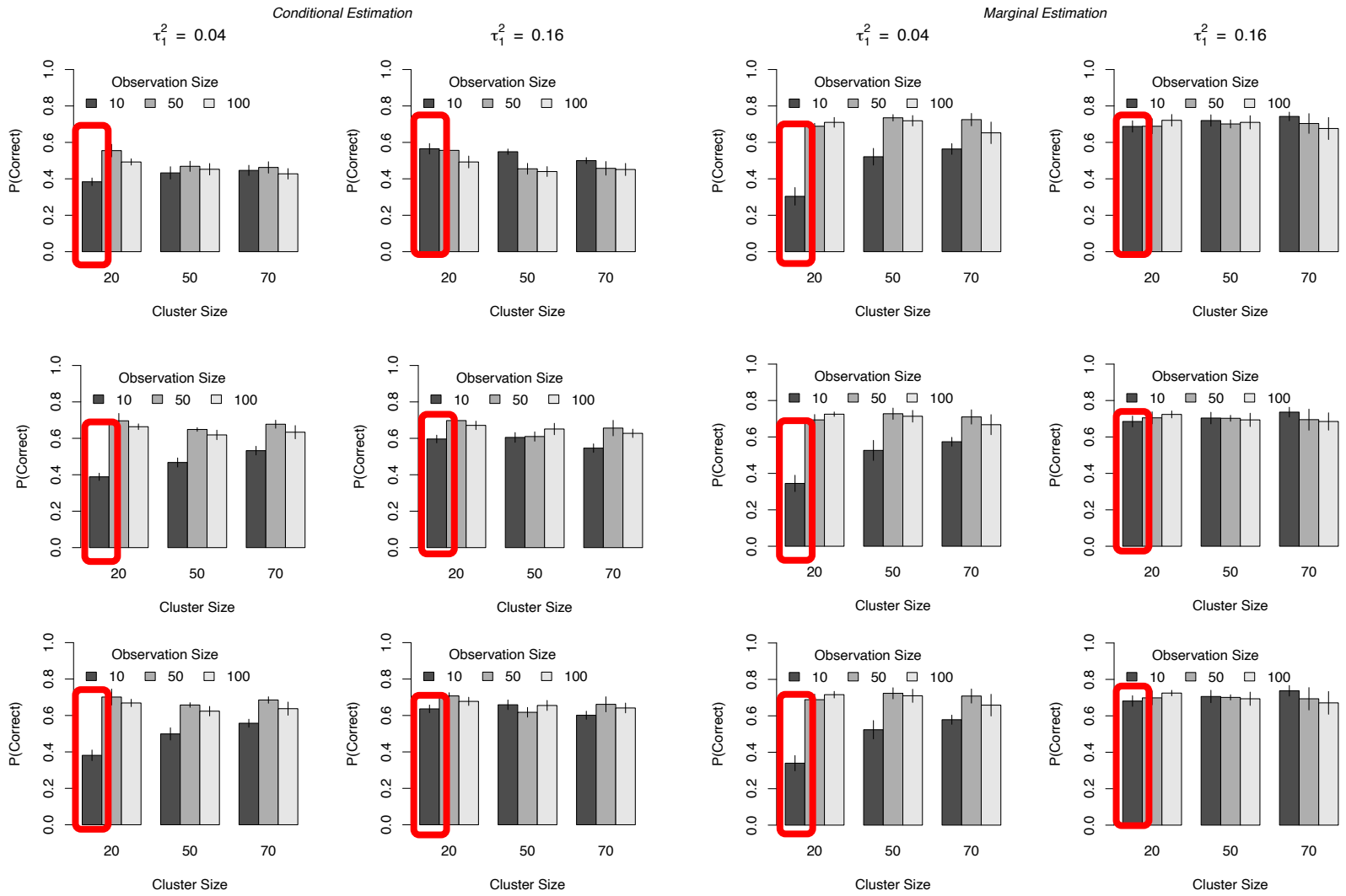
Other results

Impact of clustering variability on accuracy



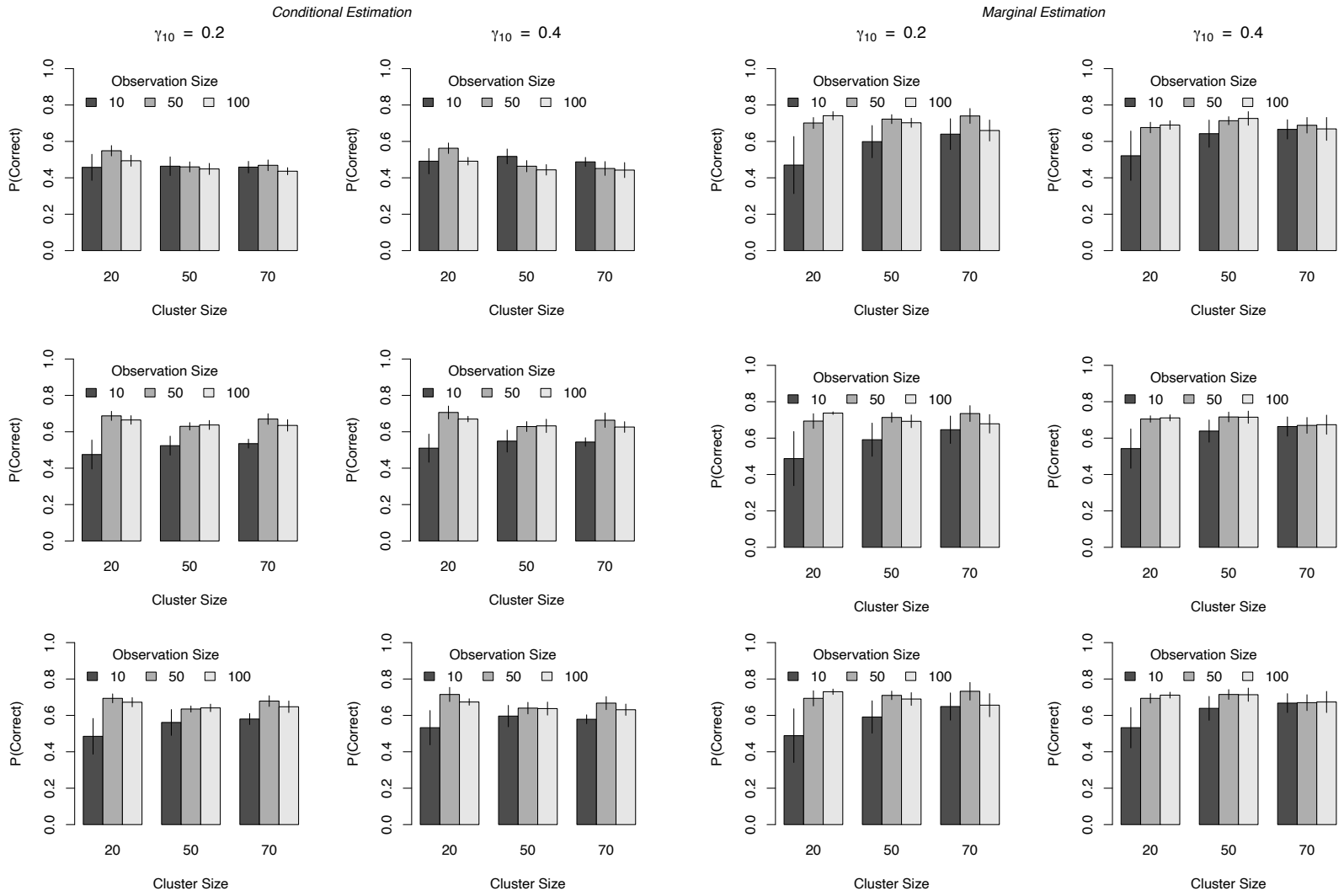
Little impact of clustering on selection accuracy, irrespective of estimation strategy

Impact of slope variability on accuracy



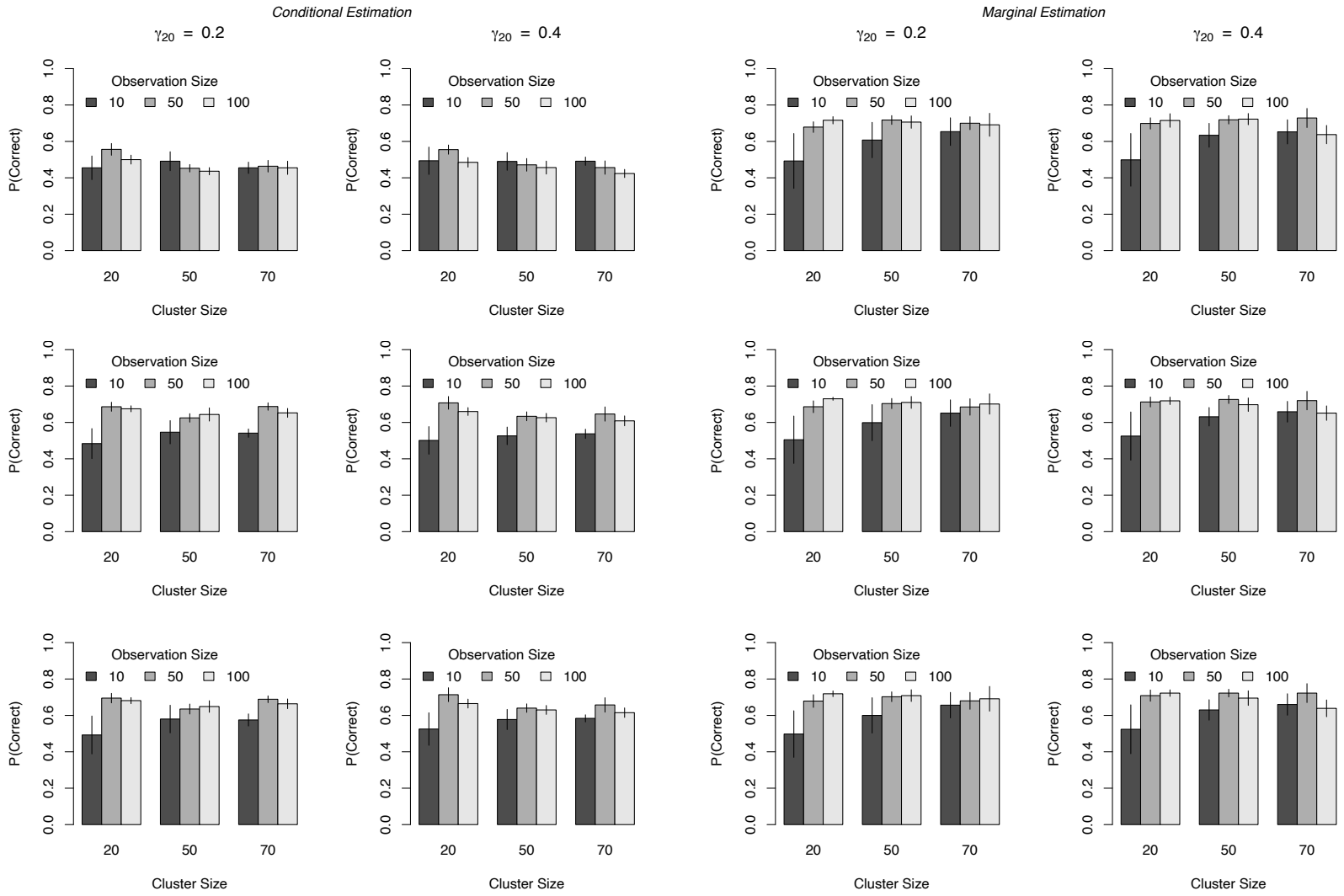
- When L1 sample size is low, higher slope variability increases model selection accuracy
- This difference is more pronounced when using marginal estimation
- Examples in **red**

Impact of X1 slope effect size on accuracy



- Little impact of X1 slope effect size on model selection accuracy, regardless of estimation strategy or IC

Impact of X1 slope effect size on accuracy



DIC

WAIC

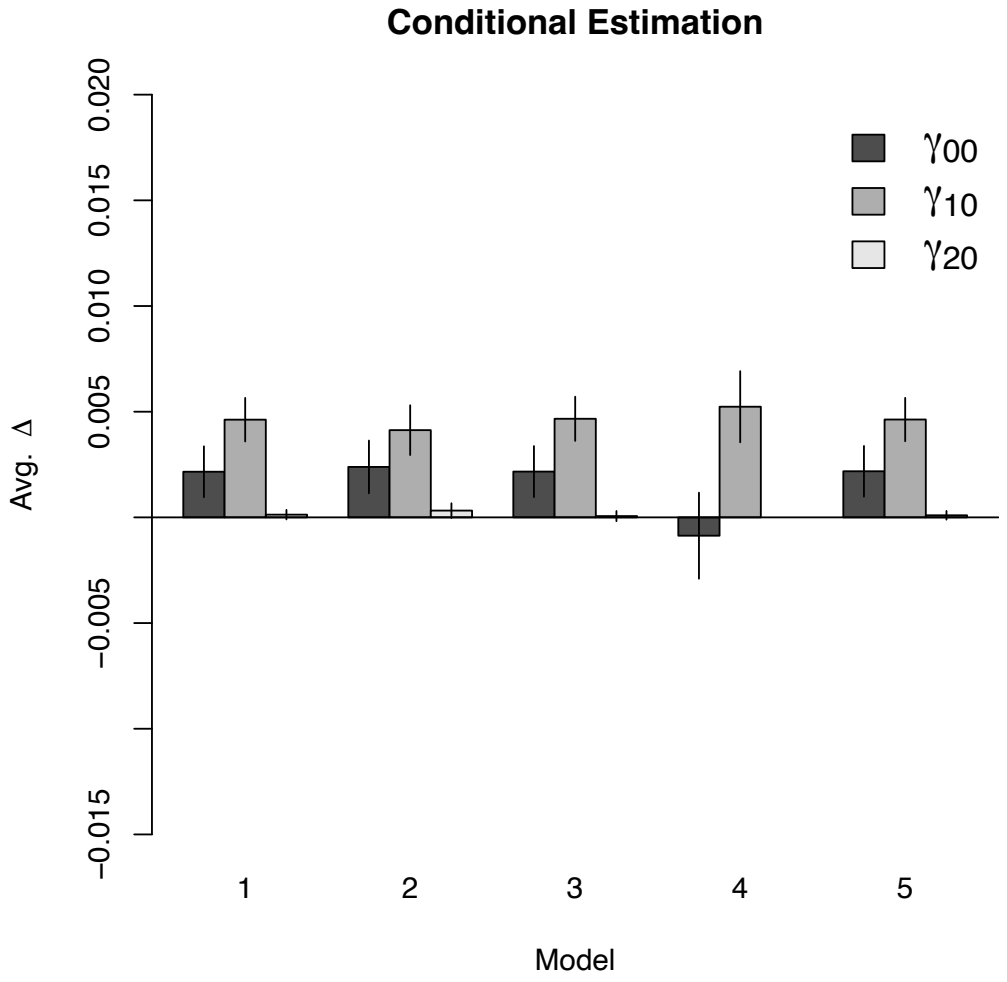
LOO

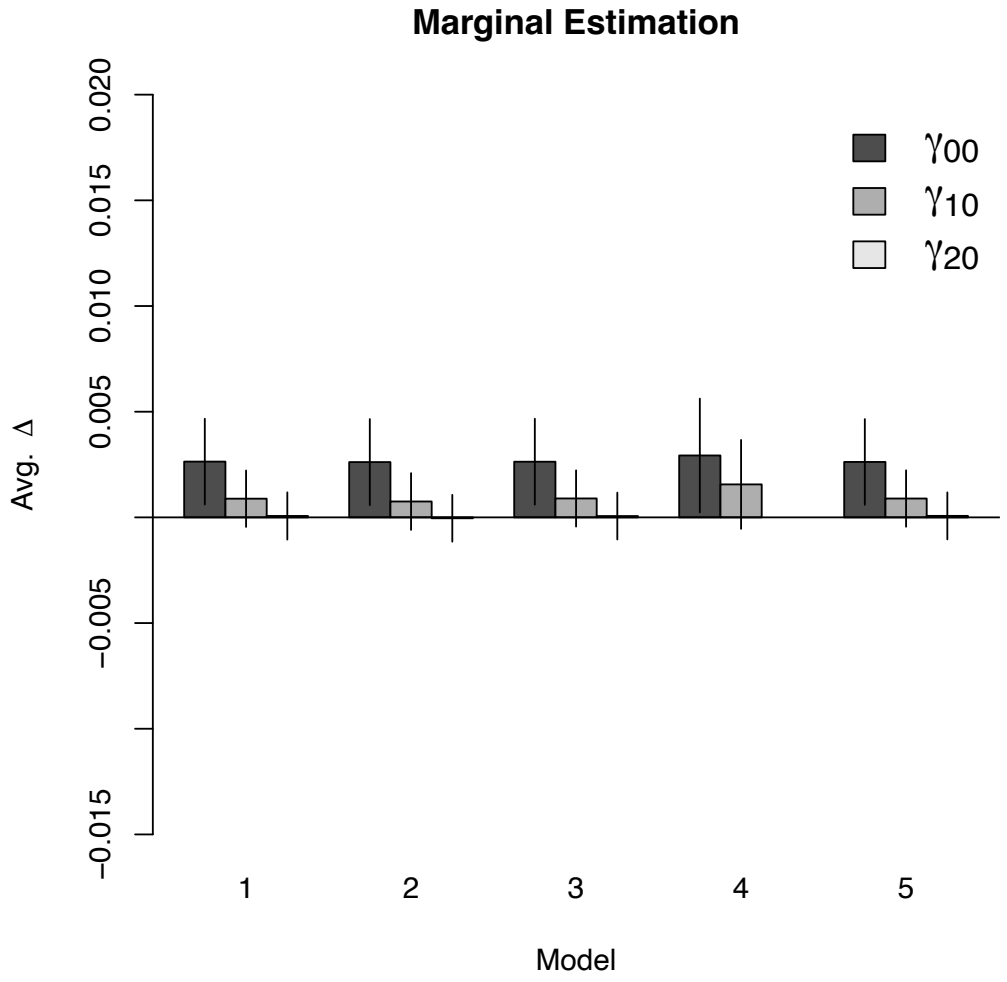
Very slight boost in accuracy for higher effect sizes, but depends on index (i.e., for LOO)

Parameter estimation bias

Fixed effects discrepancy

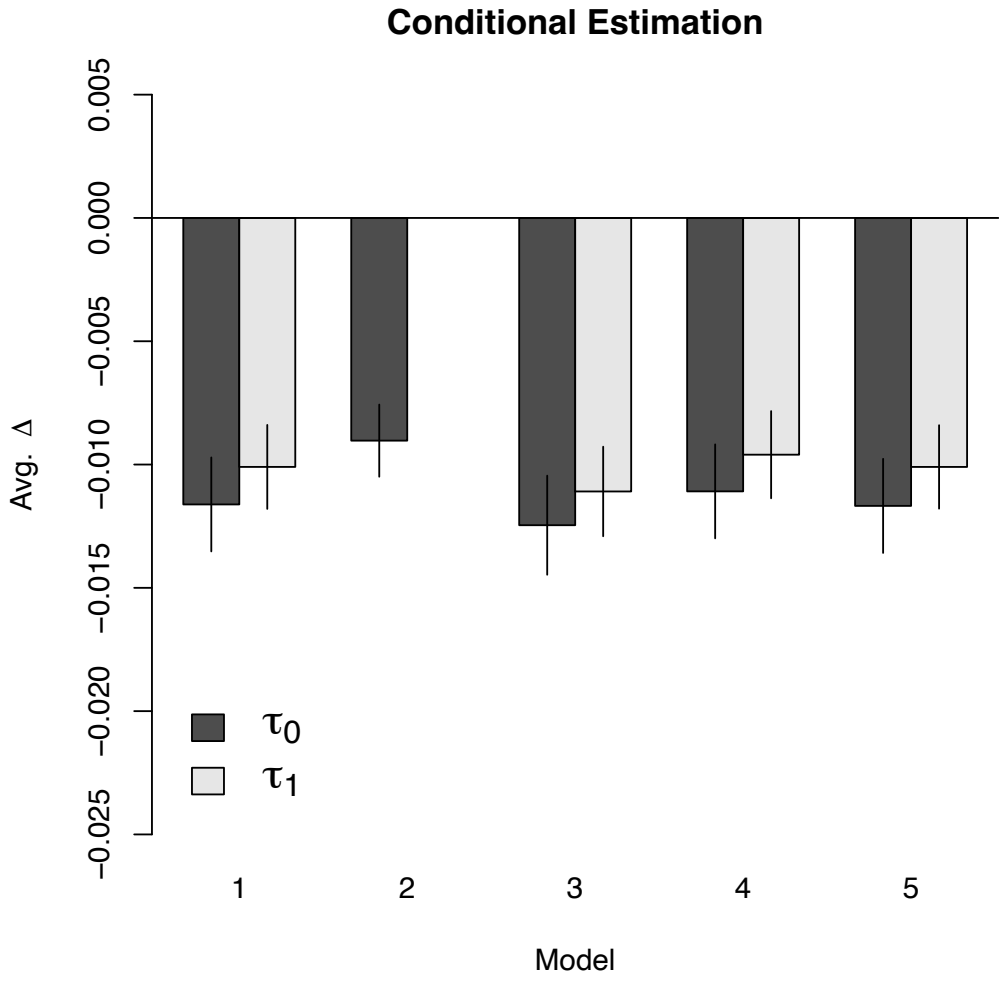
- Δ = Distance between estimated γ and true γ
- Model choice doesn't seem to radically impact fixed effect accuracy.
- Similar bias regardless of estimation strategy

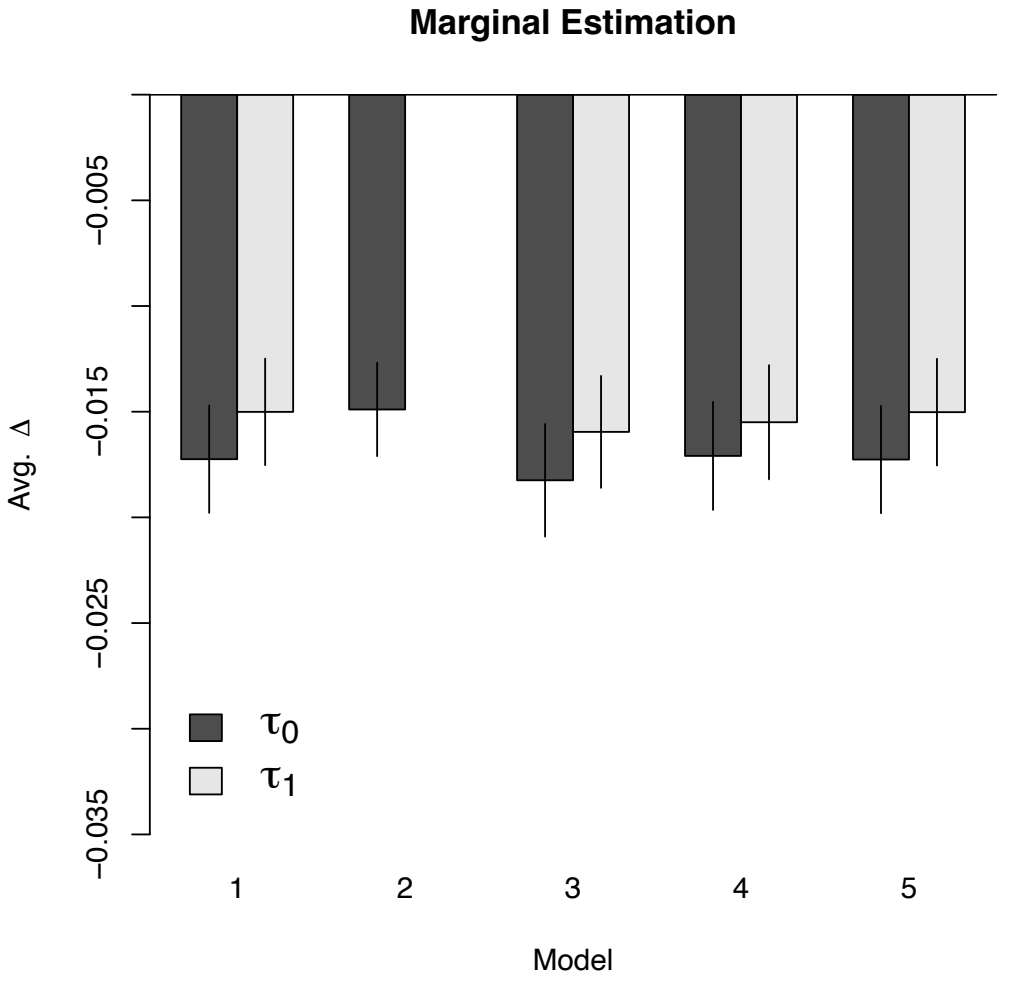




Higher ranef bias for marginal estimation

Random effects discrepancy

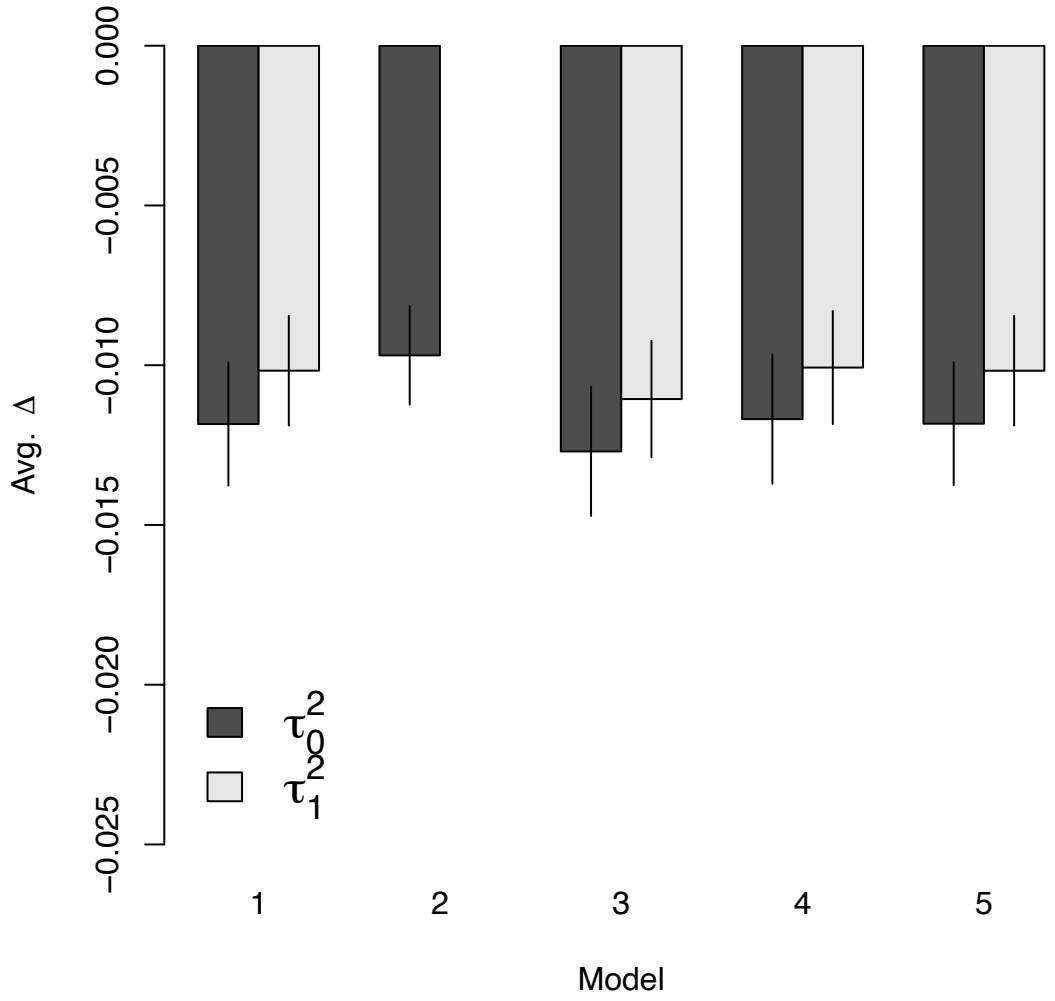




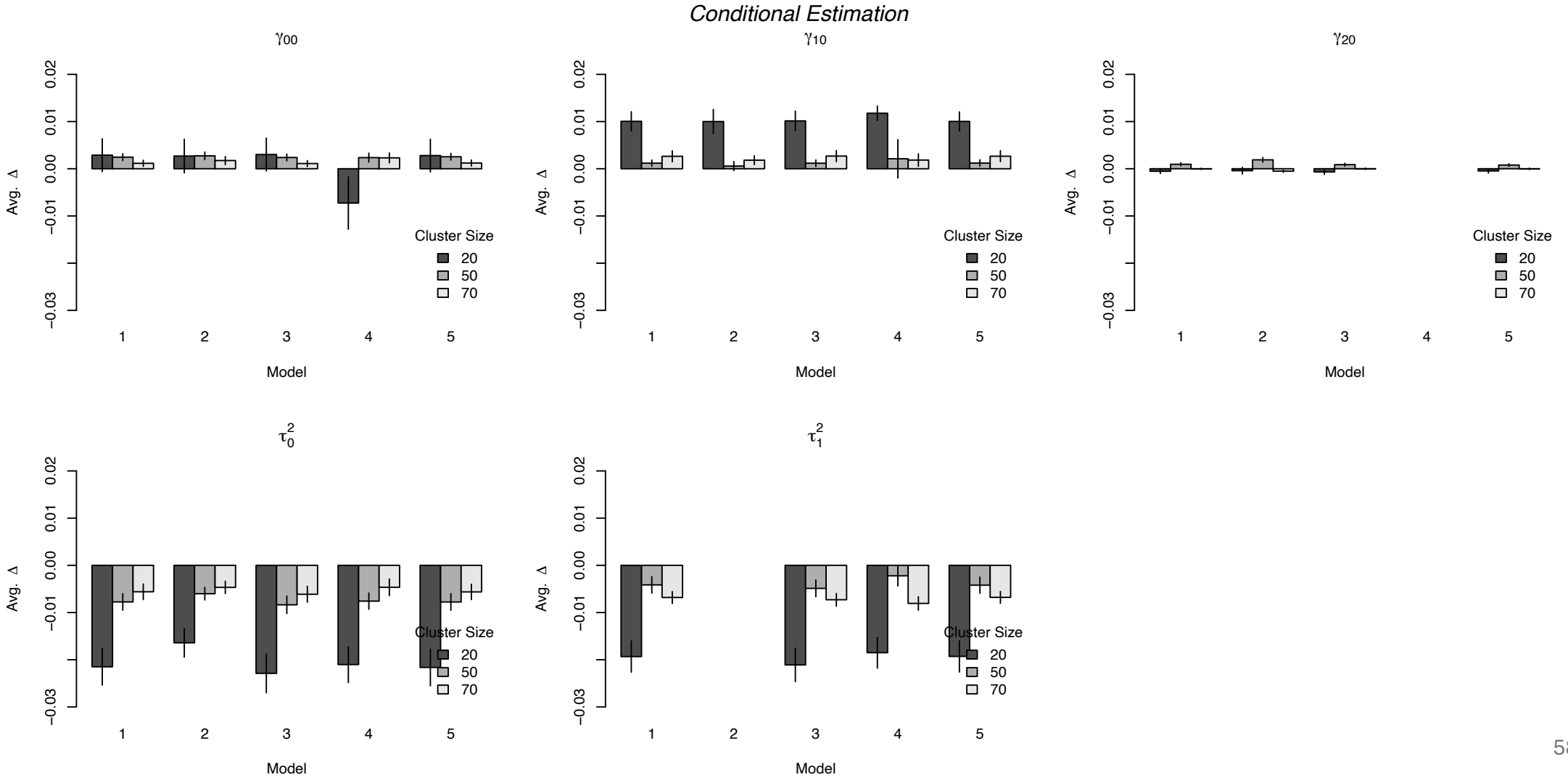
Random effects discrepancy

Δ = Distance between estimated τ and true τ

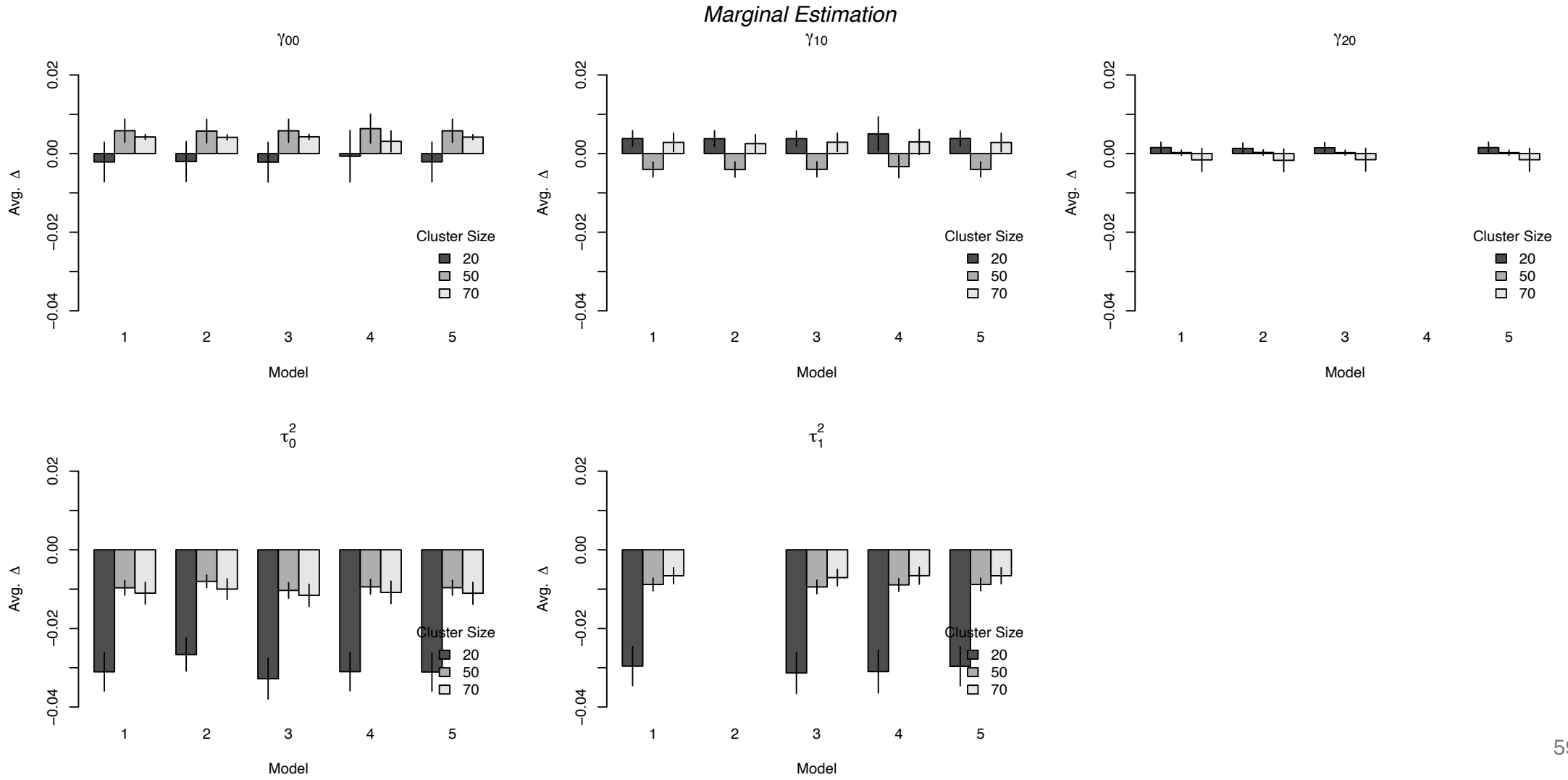
Same as with fixed effects…



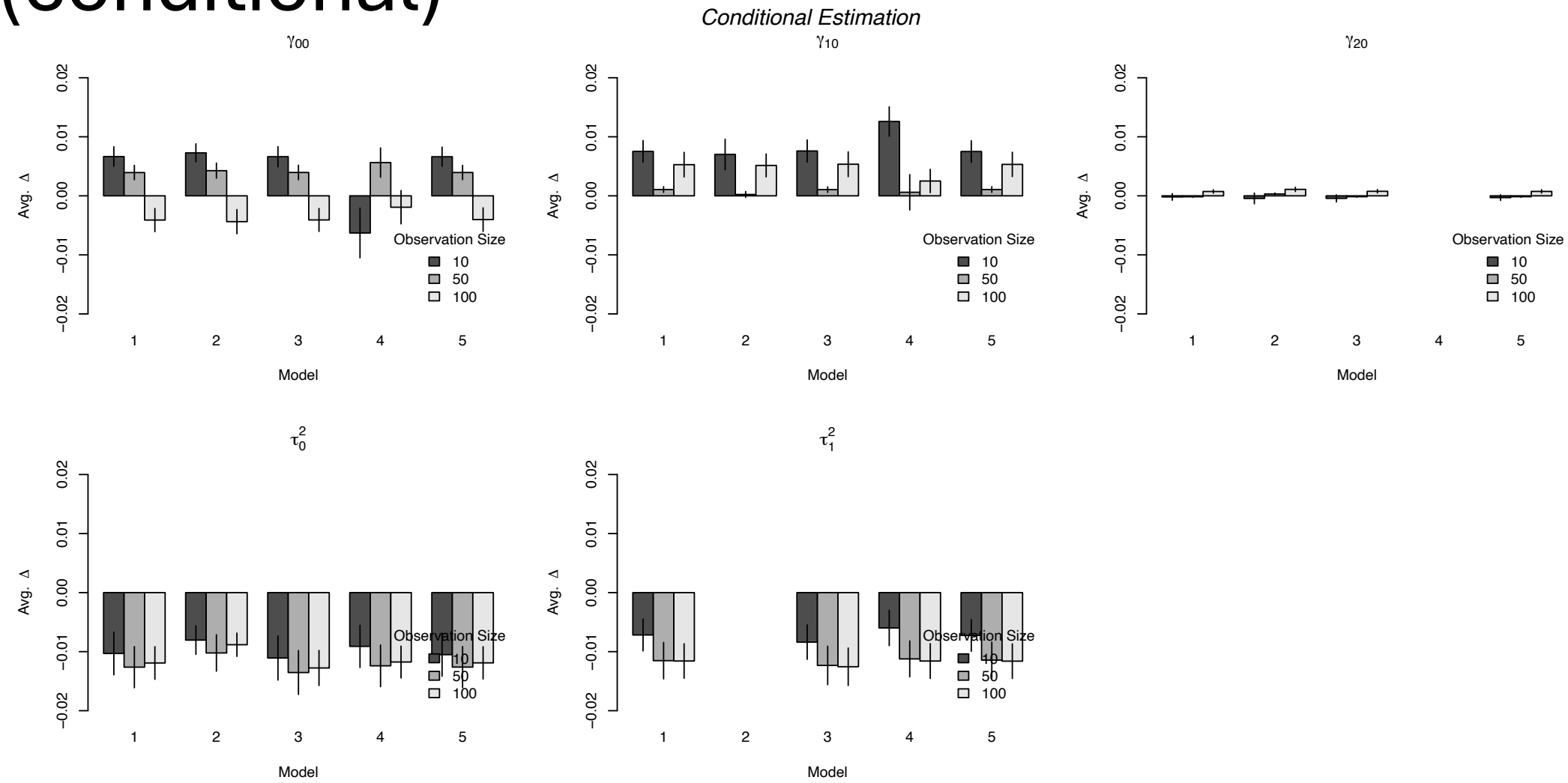
Bias by level-2 sample size (conditional)



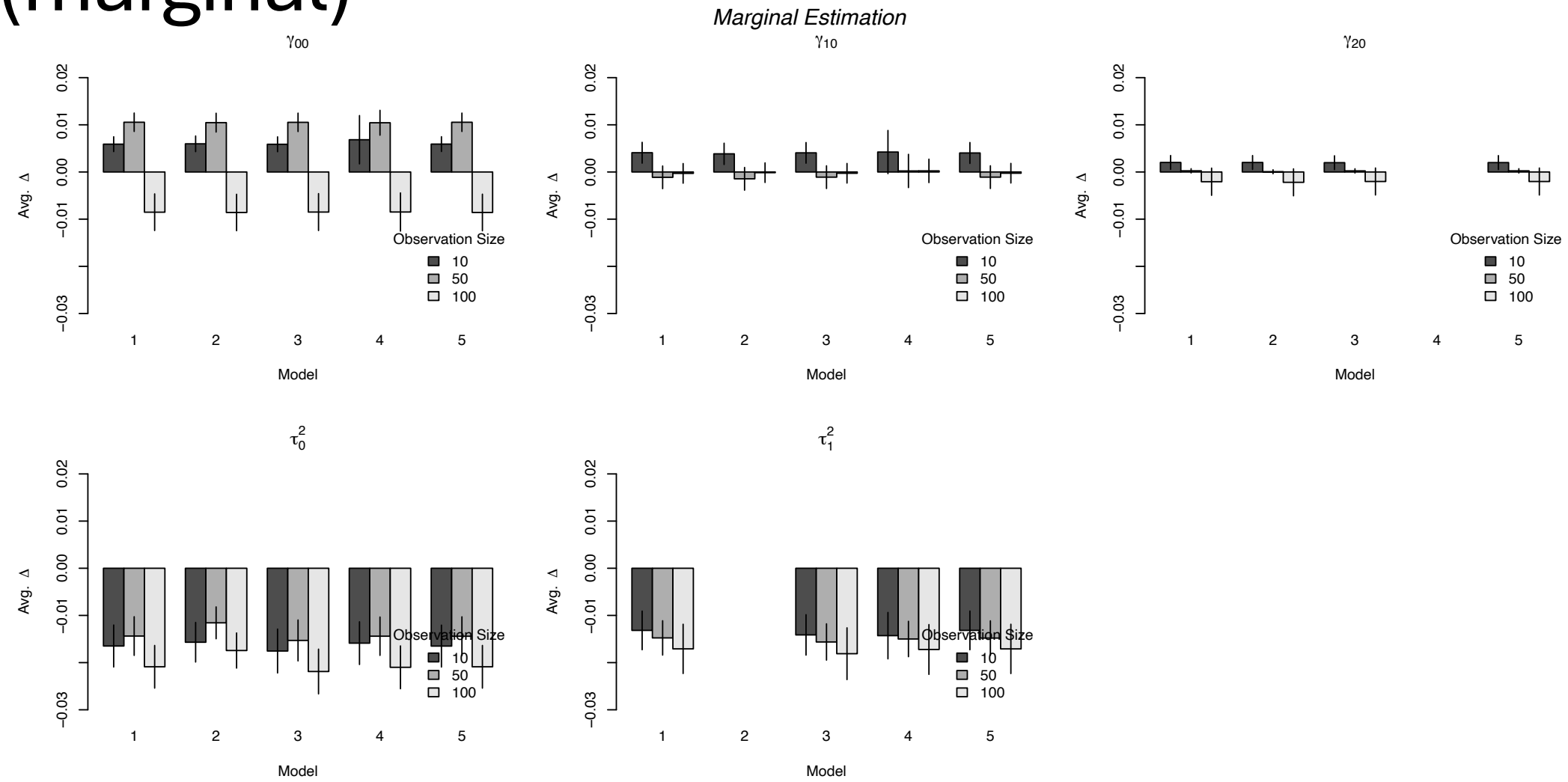
Bias by level-2 sample size (marginal)



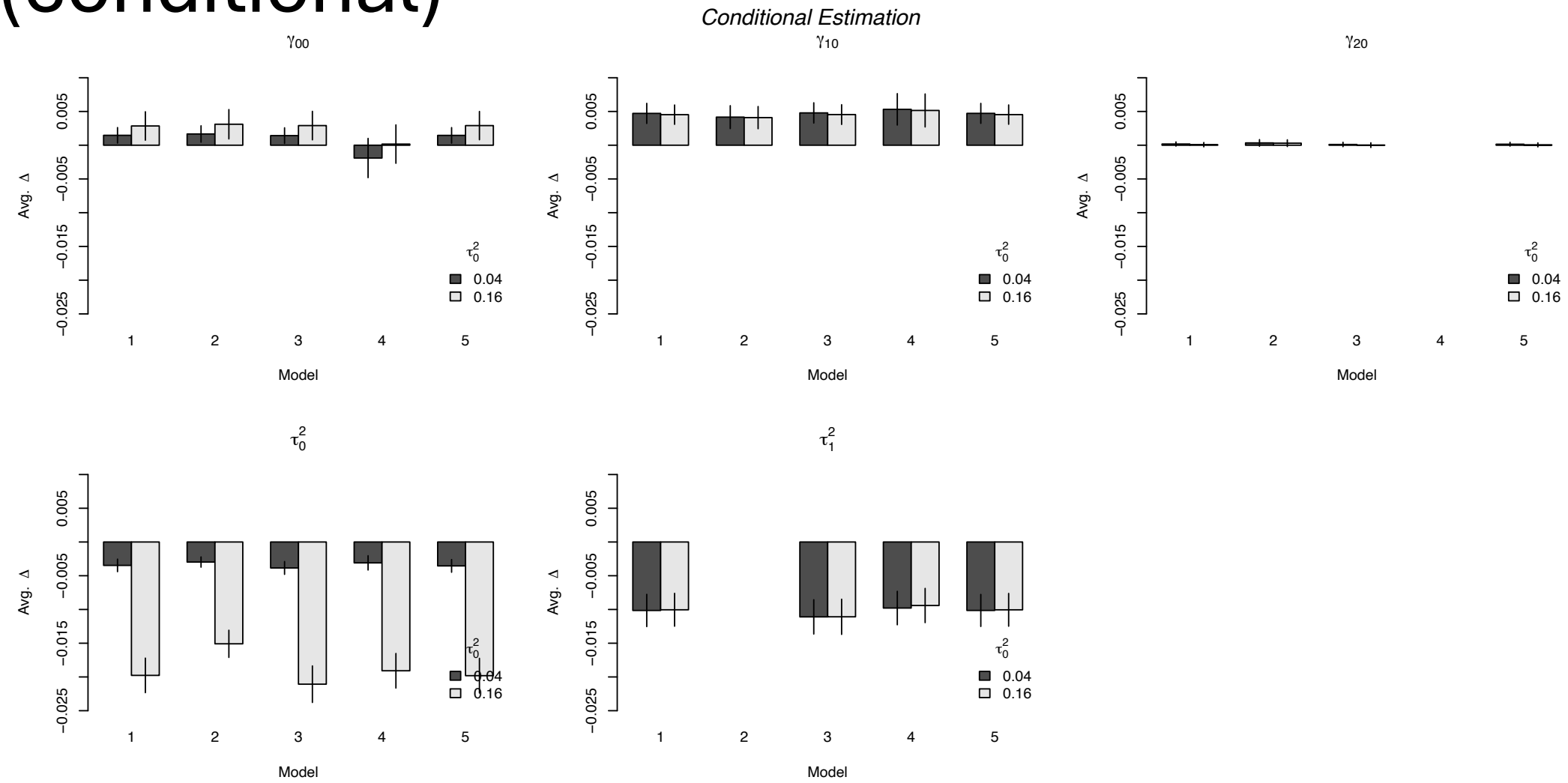
Bias by level-1 sample size (conditional)



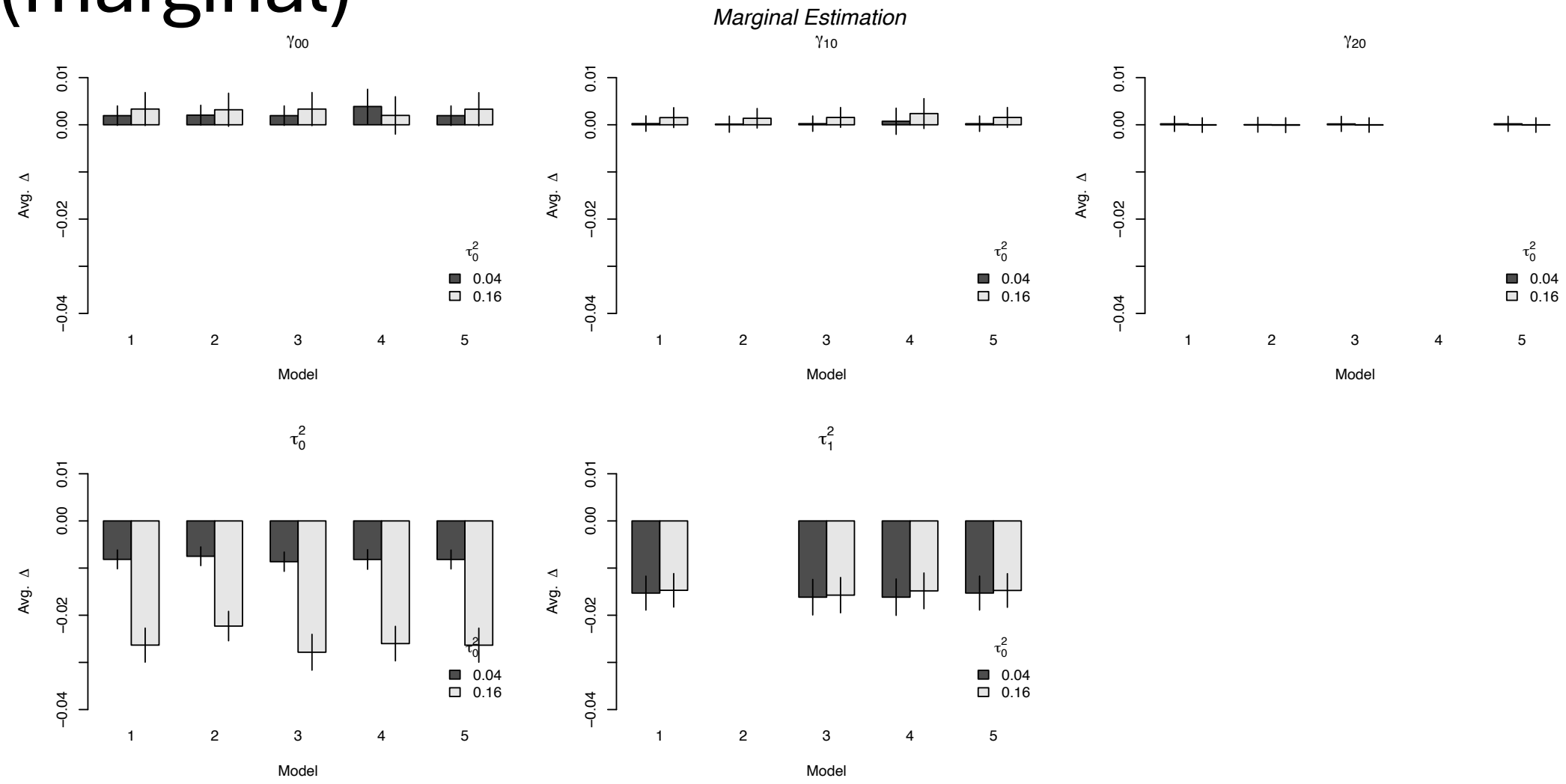
Bias by level-1 sample size (marginal)



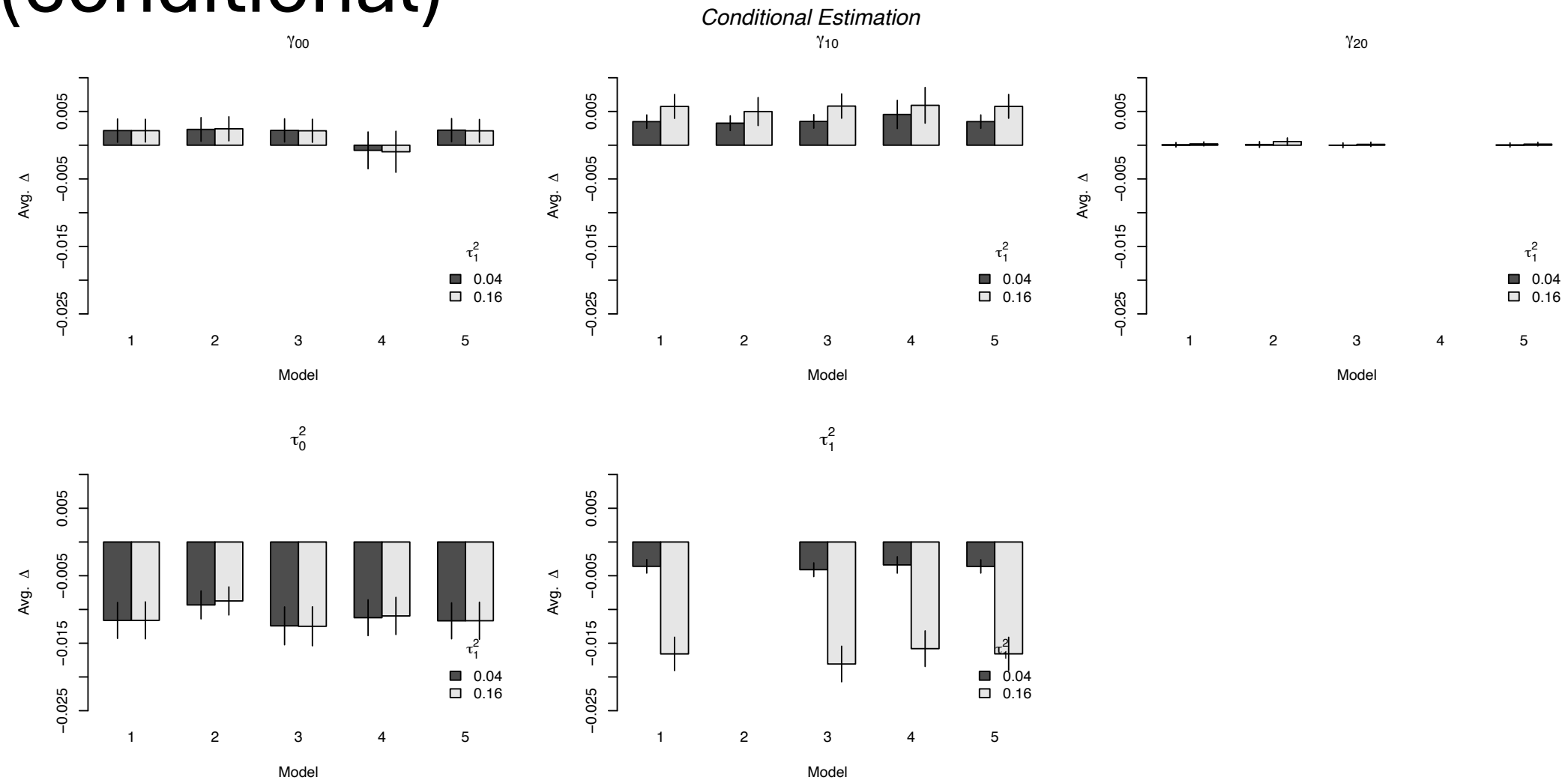
Bias by τ_0^2 magnitude (conditional)



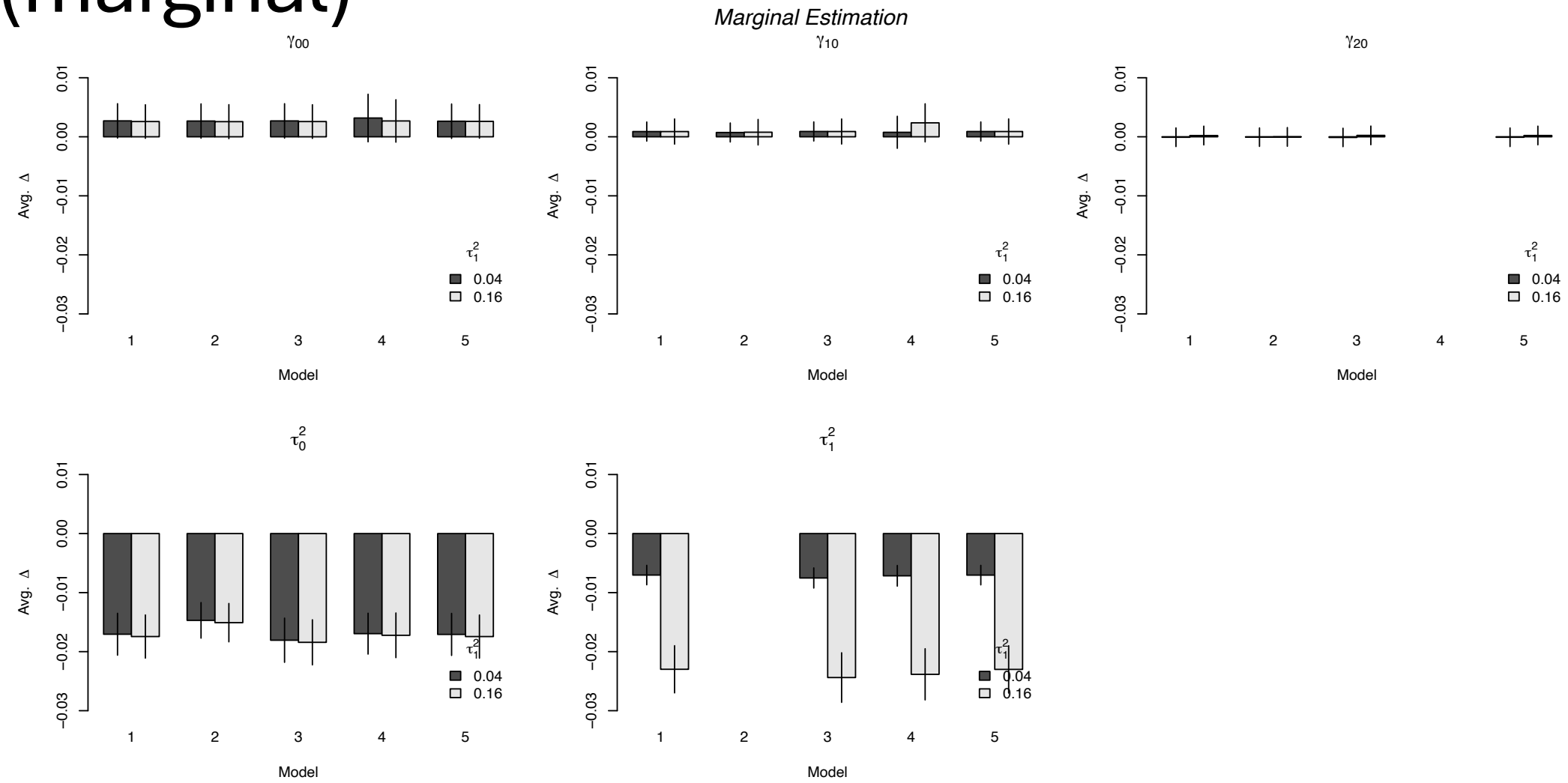
Bias by τ_0^2 magnitude (marginal)



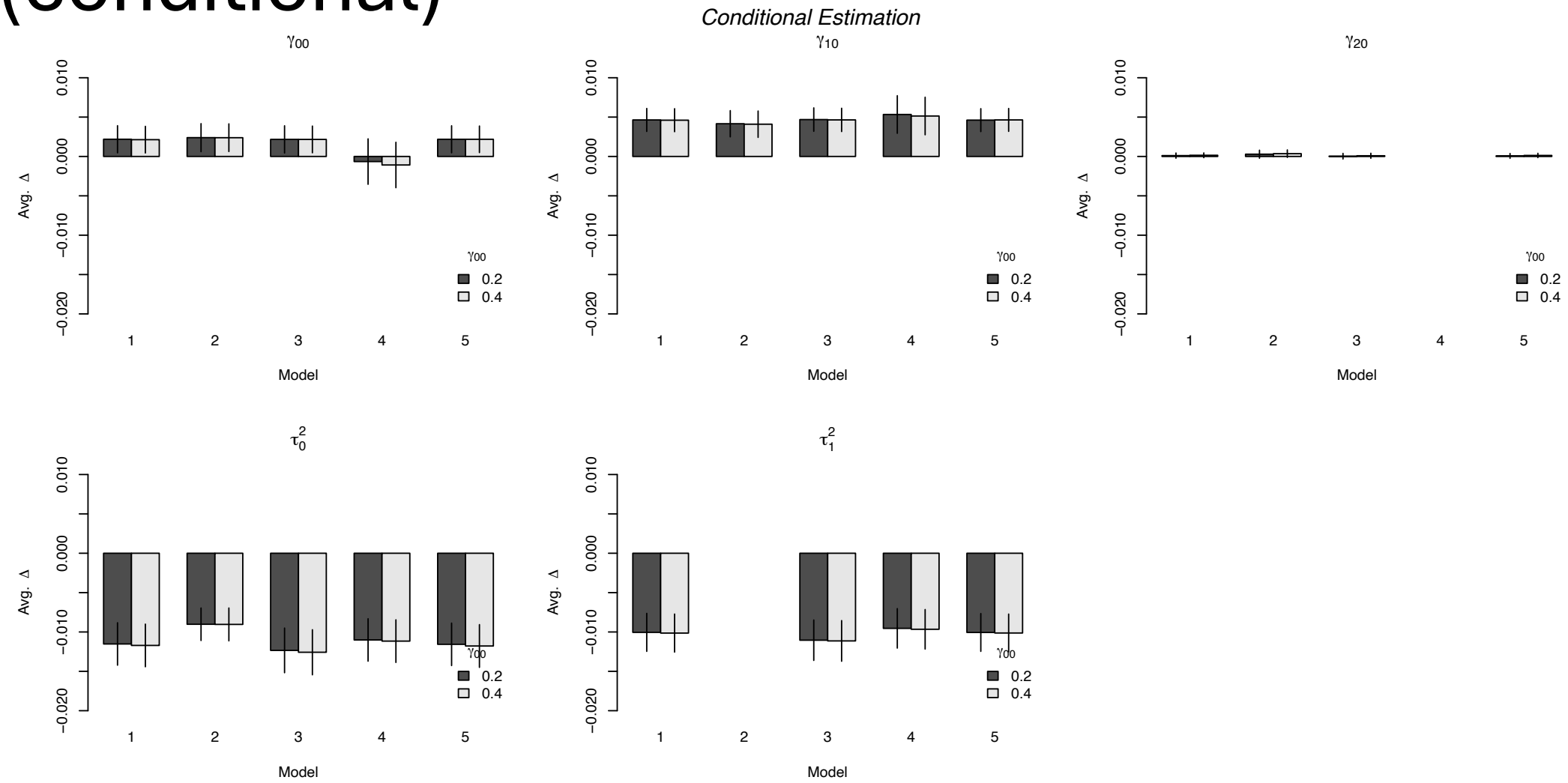
Bias by τ_1^2 magnitude (conditional)



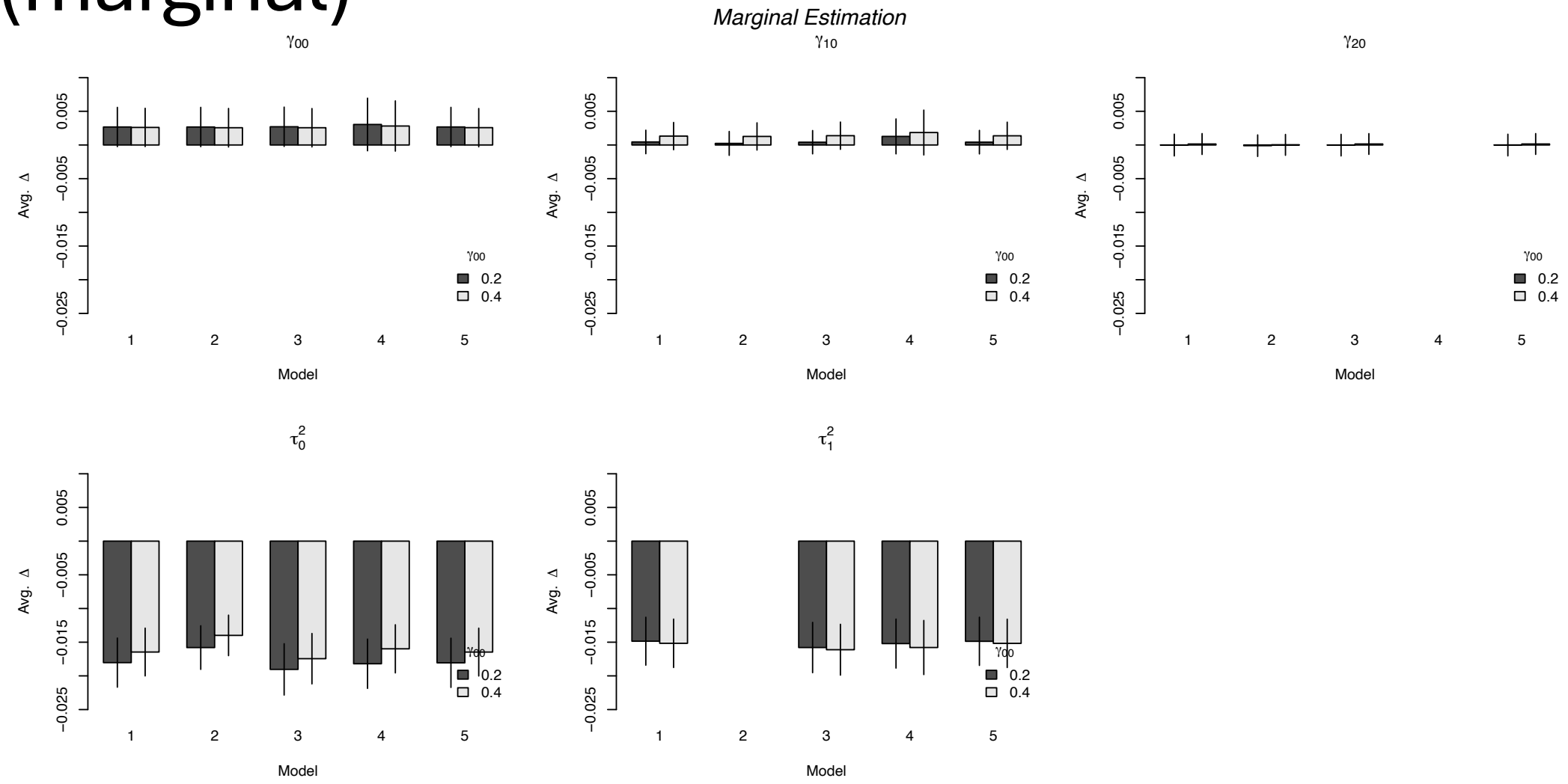
Bias by τ_1^2 magnitude (marginal)



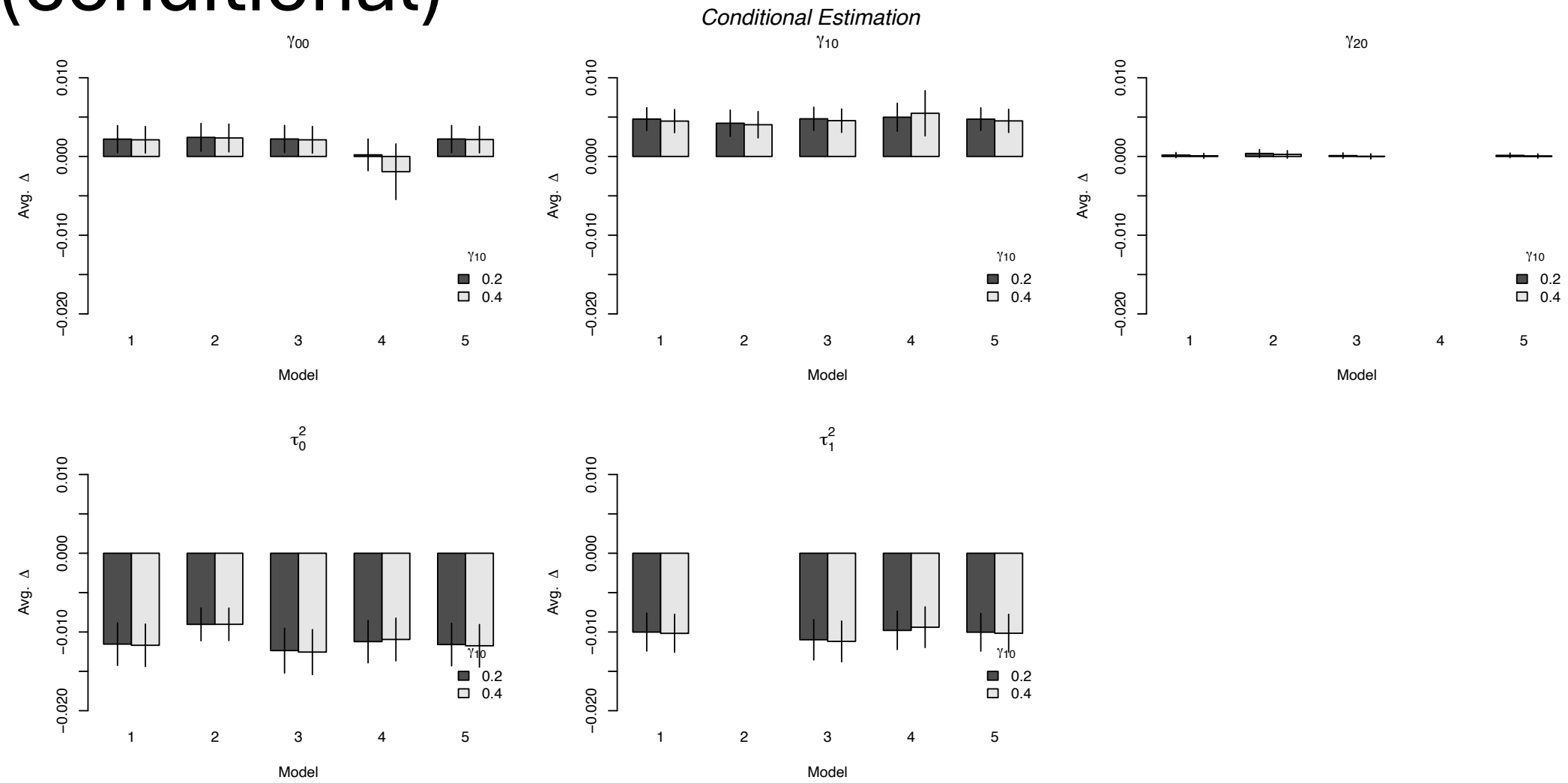
Bias by γ_{00} magnitude (conditional)



Bias by γ_{00} magnitude (marginal)



Bias by γ_{10} magnitude (conditional)



Bias by γ_{10} magnitude (marginal)