Reduction of small-sample bias of GLM parameter estimates

Annie Yao Ravleen Bajaj Samir Arora

April 3rd, 2025

Motivation

Asymptotic properties of MLE: Under usual regularity conditions, if $\hat{\boldsymbol{\theta}}$ is the MLE, then:

1. $\hat{\boldsymbol{\theta}}$ is consistent.
2. $\hat{\boldsymbol{\theta}}$ has bias of asymptotic order $\mathcal{O}(n^{-1})$, i.e., the bias vanishes as $n \to \infty$.
3. $(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx N(\mathbf{0}, F(\boldsymbol{\theta})^{-1})$, where $F(\boldsymbol{\theta})$ is the Fisher Information Matrix.

Motivation

- Do we have infinitely large sample size?
- For finite n, these properties may deteriorate, in some cases causing severe problems in inference.

Reducing Bias - A simple Recipe

- For any estimator $\hat{\boldsymbol{\theta}}$ taking values in $\boldsymbol{\Theta} \subset \mathbb{R}^p$, consider the solution of the following equation with respect to a new estimator $\tilde{\boldsymbol{\theta}}$

$$\hat{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}} = B(\boldsymbol{\theta}). \tag{1}$$

- $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - B(\boldsymbol{\theta})$ has zero bias.
- Sadly, we don't know $B(\boldsymbol{\theta})$ and usually it cannot be written in closed-form.
- All known methods to reduce bias can be usefully thought of as attempts to approximate the solution of Eq (1).
- For many common estimators, including the MLEs, the bias function can be expanded in decreasing powers of n

$$B(\boldsymbol{\theta}) = \frac{b_1(\boldsymbol{\theta})}{n} + \frac{b_2(\boldsymbol{\theta})}{n^2} + \frac{b_3(\boldsymbol{\theta})}{n^3} + \mathcal{O}(n^{-4}). \tag{2}$$

Asymptotic Bias Correction

Asymptotic Bias Correction

- One commonly used approximation of $B(\boldsymbol{\theta})$ is by $b_1(\hat{\boldsymbol{\theta}})/n$ which is the first-order bias term in Eq. (2) evaluated at $\hat{\boldsymbol{\theta}}$.
- Cox and Snell (1968) derive an expression for $b_1(\boldsymbol{\theta})/n$, when $\hat{\boldsymbol{\theta}}$ is the MLE.

$$\frac{b_1(\boldsymbol{\theta})}{n} = -\{F(\boldsymbol{\theta})\}^{-1} A(\boldsymbol{\theta}), \tag{3}$$

where, $A(\boldsymbol{\theta})$ is a p-dimensional vector with components

$$A_t(\boldsymbol{\theta}) = \frac{1}{2} tr[\{F(\boldsymbol{\theta})\}^{-1}\{P_t(\boldsymbol{\theta}) + Q_t(\boldsymbol{\theta})\}],$$

$$P_t(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}[U(\boldsymbol{\theta})U^T(\boldsymbol{\theta})U_t(\boldsymbol{\theta})],$$

$$Q_t(\boldsymbol{\theta}) = -E_{\boldsymbol{\theta}}[I(\boldsymbol{\theta})U_t(\boldsymbol{\theta})], \quad (t = 1, \cdots p).$$

where, $I(\boldsymbol{\theta})$ is the observed information matrix and $U(\boldsymbol{\theta})$ is the score function.

Asymptotic Bias Correction

- So the new bias-adjusted estimator is

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} - b_1(\hat{\boldsymbol{\theta}})/n.$$

- Efron (1975) showed that:
  1. $\tilde{\boldsymbol{\theta}}$ has bias of order $o(n^{-1})$, which is of smaller order than the $\mathcal{O}(n^{-1})$ bias of the MLE.
  2. The asymptotic variance of any estimator with $\mathcal{O}(n^{-2})$ bias is greater or equal to the asymptotic variance of $\tilde{\boldsymbol{\theta}}$, i.e., **second order efficiency**.

Asymptotic Bias Correction in GLMs

From the general expression given in Eq. (3), Cordeiro and McCullagh (1991) derived the expression for first-order bias of coefficients of a linear predictor in GLM.

$$b_1(\hat{\beta})/n = -(2\phi)^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}_d\mathbf{F}\mathbf{1} \tag{4}$$

where,

$$\mathbf{Z} = \{z_{ij}\} = \mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T, \quad \mathbf{Z}_d = diag\{z_{11}, \cdots, z_{nn}\},$$

$$F = diag\{f_{11}, \cdots f_{nn}\}, \quad f = V^{-1}(d\mu/d\eta)(d^2\mu/d\eta^2),$$

and $\mathbf{1}$ is an $n \times 1$ vector of ones.

Using Asymptotic Bias Correction

```r
# Installing the required libraries
install.packages("brglm2")

# Loading the required packages
library(brglm2)
library(tidyverse)

# Loading and cleaning the dataset
school_binary <- read.table("C:/Users/samir/Downloads/school.txt", header=T)
school_binomial <- school_binary %>%
  group_by(Grade, Sex, Participate) %>%
  summarize(n=n(), y=sum(Pass))
school_binomial$Grade <- factor(school_binomial$Grade)

# Fitting the usual model
fit_binomial <- glm(cbind(y,n-y) ~ Grade + Sex + Participate,
                    binomial(logit),
                    data = school_binomial)
summary(fit_binomial)

# Fitting the bias reduced version
fit_binomial_bias_corrected <- glm(cbind(y,n-y) ~ Grade + Sex + Participate,
                                   binomial(logit),
                                   data = school_binomial,
                                   method=brglm_fit, type = "correction")
summary(fit_binomial_bias_corrected)
```

Figure: Using Cordeiro and McCullagh (1991) bias correction in practice

Using Asymptotic Bias Correction

```
> summary(fit_binomial)

Call:
glm(formula = cbind(y, n - y) ~ Grade + Sex + Participate, family = binomial(logit),
    data = school_binomial)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.4013     0.3374   4.153 3.28e-05 ***
Grade8        0.3622     0.3493   1.037    0.300
Grade10      -0.4914     0.3202  -1.535    0.125
Grade12      -0.7879     0.3207  -2.457    0.014 *
Sex          -0.9308     0.2160  -4.310 1.63e-05 ***
Participate   0.1739     0.2266   0.768    0.443
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 47.137  on 15  degrees of freedom
Residual deviance: 11.513  on 10  degrees of freedom
AIC: 77.899

Number of Fisher Scoring iterations: 4
```

Figure: Usual GLM summary

Using Asymptotic Bias Correction

```
> summary(fit_binomial_bias_corrected)

Call:
glm(formula = cbind(y, n - y) ~ Grade + Sex + Participate, family = binomial(logit),
    data = school_binomial, method = brglm_fit, type = "correction")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5331  -0.4312  -0.0655   0.6510   1.4830

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.3749     0.3359   4.093 4.25e-05 ***
Grade8        0.3572     0.3474   1.028   0.3038
Grade10      -0.4794     0.3191  -1.503   0.1330
Grade12      -0.7720     0.3197  -2.415   0.0157 *
Sex          -0.9164     0.2152  -4.258 2.06e-05 ***
Participate   0.1735     0.2259   0.768   0.4426
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 47.137  on 15  degrees of freedom
Residual deviance: 11.534  on 10  degrees of freedom
AIC:  77.92

Type of estimator: correction (bias correction)
Number of Fisher Scoring iterations: 1
```

Figure: GLM with Cordeiro and McCullagh (1991) bias correction summary

Trade-offs in Asymptotic Bias Correction

Advantages

- Simplicity

Disadvantages

- Inherit any of the instabilities of the original estimator. What if MLE is not finite?
- Only applicable when $b_1(\theta)/n$ is available in closed form.

Firth Method

Introduction

In a regular model with a p-dimensional parameter θ, the asymptotic bias of the maximum likelihood estimate $\hat{\theta}$ may be written as

$$B(\theta) = \frac{b_1(\theta)}{n} + \frac{b_2(\theta)}{n^2} + \ldots,$$

where n is usually interpreted as the number of observations unless otherwise stated.

FOCUS: A *general* method for reducing bias, with a specific aim being the removal of the term $\mathcal{O}(n^{-1})$

Previously Studied Approach

A standard approach simply substitutes $\hat{\theta}$ for the unknown θ in $b_1(\theta)/n$; the bias-corrected estimate is then calculated as

$$\tilde{\theta} = \hat{\theta} - \frac{b_1(\hat{\theta})}{n}.$$

These methods succeed in removing the term $\frac{b_1(\theta)}{n}$ from the asymptotic bias.

A common feature of the previous approaches is that they are 'corrective', rather than 'preventive' in character.

Motivation for Firth's Method

Firth's method is a procedure for reducing the leading-order (i.e., $\mathcal{O}(1/n)$) bias of maximum likelihood estimators. Rather than computing the ordinary MLE and then adjusting it *after* the fact, Firth's approach *modifies* the score equations themselves (or equivalently, penalizes the log-likelihood).

A general modification of the score function is of the form:

$$U^*(\theta) = U(\theta) + A(\theta)$$

Firth's Method

We make a modification to $U(\boldsymbol{\theta})$:

$$U^*(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\frac{B_1(\boldsymbol{\theta})}{n}$$

and we set

$$U^*(\theta) = 0$$

Thus, the bias equation can be re-written as:

$$\{F(\theta)^{-1}\}U(\theta) = \frac{B_1(\theta)}{n}$$

Firth's Method

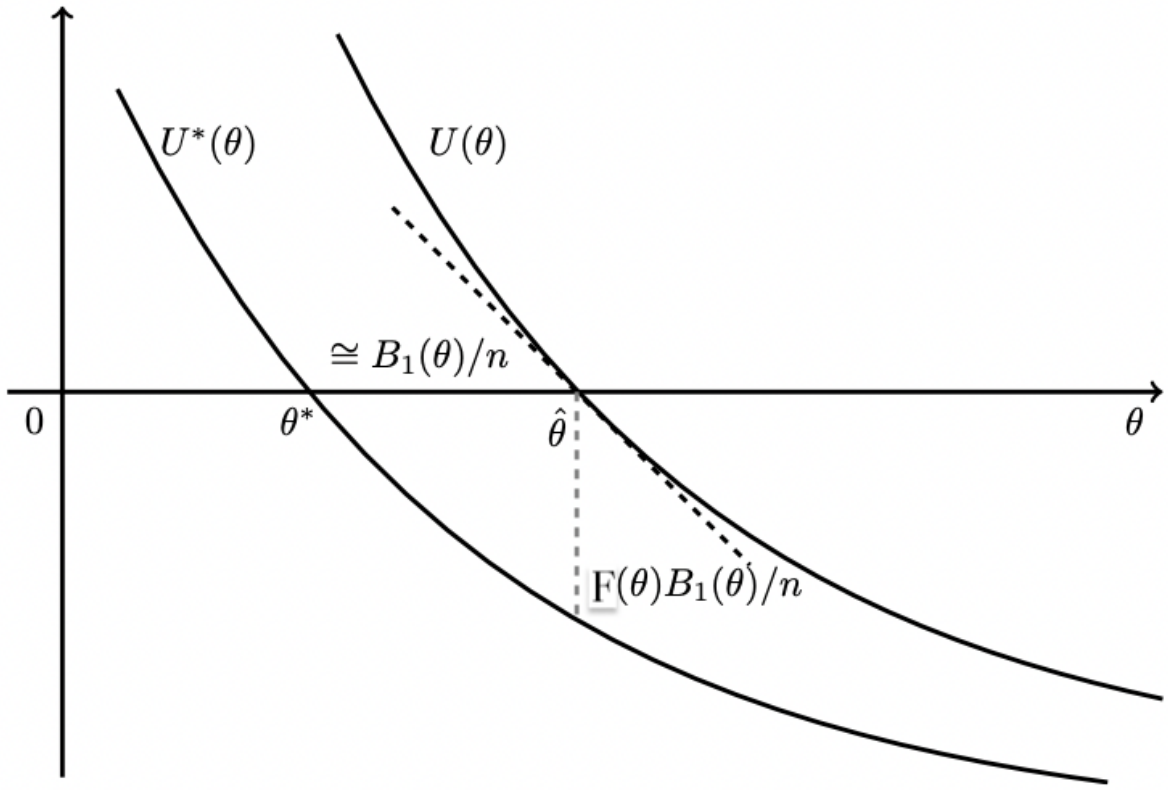


Figure 1: Modification of Score Function

where $F(\theta)$ denotes Fisher's information of the sample, defined as the negative expected value of the first derivative of $U(\theta)$. The modified score function $U^*(\theta)$ originates from the simple triangle geometry shown in Figure 1 adapted from Firth (1993).

Firth's Method

If the MLE $\hat{\theta}$ has a positive first-order bias of $B_1(\theta)/n$, it can be removed by shifting the score function downward by $F(\theta)B_1(\theta)/n$, where the gradient is $U'(\theta) = -F(\theta)$

$$\frac{\partial U(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = -F(\boldsymbol{\theta}).$$

For the canonical parameter of an exponential model, we remove the $\mathcal{O}(n^{-1})$ bias term.

Then, Firth's Adjusted Likelihood is:

$$\ell^*(\theta) = \ell(\theta) + \frac{1}{2}\log|\mathbf{F}(\theta)|$$

Jeffreys Prior as a Bias Reducing Function

The solution of

$$U^*(\theta) \equiv U(\theta) + A(\theta) = 0$$

locates a stationary point of:

$$\ell^*(\theta) = \ell(\theta) + \frac{1}{2} \log |\mathbf{F}(\theta)|$$

or equivalently

$$\mathcal{L}^*(\theta) = \mathcal{L}|\mathbf{F}(\theta)|^{1/2}$$

The penalty function $|\mathbf{F}(\theta)|^{1/2}$ is the Jefferys invariant prior.

The Logistic Regression Model

For *ith* subject, the response y_i is Binomial distributed with probability of success π_i where $x_i = (x_1 \cdots x_p)$

$$Y_i \sim Bin(n_i, \pi_i)$$

$$log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

and

$$\hat{\pi}_i = \frac{e^{\beta_0 + \mathbf{x_i^T}\beta}}{1 + e^{\beta_0 + \mathbf{x_i^T}\beta}}$$

where β_0 is an intercept term and $\beta = (\beta_1 \cdots \beta_p)^T$

The Logistic Regression Model

Let $\theta = (\beta_0, \beta)^T$

The likelihood is:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}$$

The log-likelihood is:

$$\ell(\theta) = log\mathcal{L} = \sum_{i=1}^{n} \{y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i)\} .$$

Firth's Logistic Regression Model

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = U(\boldsymbol{\theta}) = \sum_{i=1}^{n} (y_i - \pi_i)\boldsymbol{x}_i.$$

We make a modification to $U(\boldsymbol{\theta})$ such that:

$$U^*(\boldsymbol{\theta}) = U(\boldsymbol{\theta}) - F(\boldsymbol{\theta})\frac{B_1(\boldsymbol{\theta})}{n}$$

and hence a modified estimate θ^* is a given as a solution $U^*(\theta) = 0$.

Firth's Logistic Regression Model

In case of logistic regression:

$$\mathbf{F}(\theta) = -\frac{\partial^2 \ell(\theta)}{\partial \theta \partial \theta'} = \mathbf{X}^T \mathbf{W} \mathbf{X},$$

where, $\mathbf{X}$ is an $n \times (p+1)$ design matrix with elements in the first column being 1, and $\mathbf{W}$ is an $n \times n$ diagonal matrix with general element $\pi_i(1-\pi_i)$.

Data

```
##   y           X1
## 1 0  1.24151347
## 2 0  0.07128281
## 3 1  1.54630897
## 4 1  2.54497817
## 5 1  0.74195087
## 6 0 -0.32570867
```

R Tutorial

```
##
## Call:
## glm(formula = y ~ X1, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1188     0.3069   0.387   0.6988
## X1           -0.4180     0.1755  -2.382   0.0172 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 62.379  on 48  degrees of freedom
## AIC: 66.379
##
## Number of Fisher Scoring iterations: 4
```

R Tutorial

```
##
## Call:
## glm(formula = y ~ X1, family = binomial(link = "logit"), data = data,
##     method = brglmFit, type = "AS_mean")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.53927  -1.05216   0.08108   1.01932   1.75341
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.1106     0.3042   0.364    0.716
## X1           -0.3862     0.1711  -2.258    0.024 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 69.315  on 49  degrees of freedom
## Residual deviance: 62.412  on 48  degrees of freedom
## AIC:  66.412
##
## Type of estimator: AS_mean (mean bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 3
```

Log-F Prior for Logistic Regression

Greenland & Mansournia (2015)

Single-Parameter Setting

Consider a simple single-parameter model:

$$\log\Big(\frac{\pi}{1-\pi}\Big) = \beta_0$$

The likelihood function for β_0 is

$$L(\beta_0) \;=\; \pi^y\,(1-\pi)^{n-y},$$

This has a closed-form maximum at

$$\hat{\beta}_0 \;=\; \log\Big(\tfrac{y}{n-y}\Big).$$

Single-Parameter with Jeffrey's Prior

The penalized likelihood under Firth's method with Jeffrey's Prior:

$$L^*(\beta_0) \propto \; L(\beta_0) * F(\beta_0)^{1/2}$$

Where the Fisher's Information in this setting is:

$$F(\beta_0) \; = \; \mathbb{E}\Big[-\frac{d^2}{d\beta_0^2}\,\ell(\beta_0)\Big] = \frac{n\,e^{\beta_0}}{\left(1+e^{\beta_0}\right)^2}$$

log-$F(m, m)$ Prior

Instead of Jeffrey's Prior:

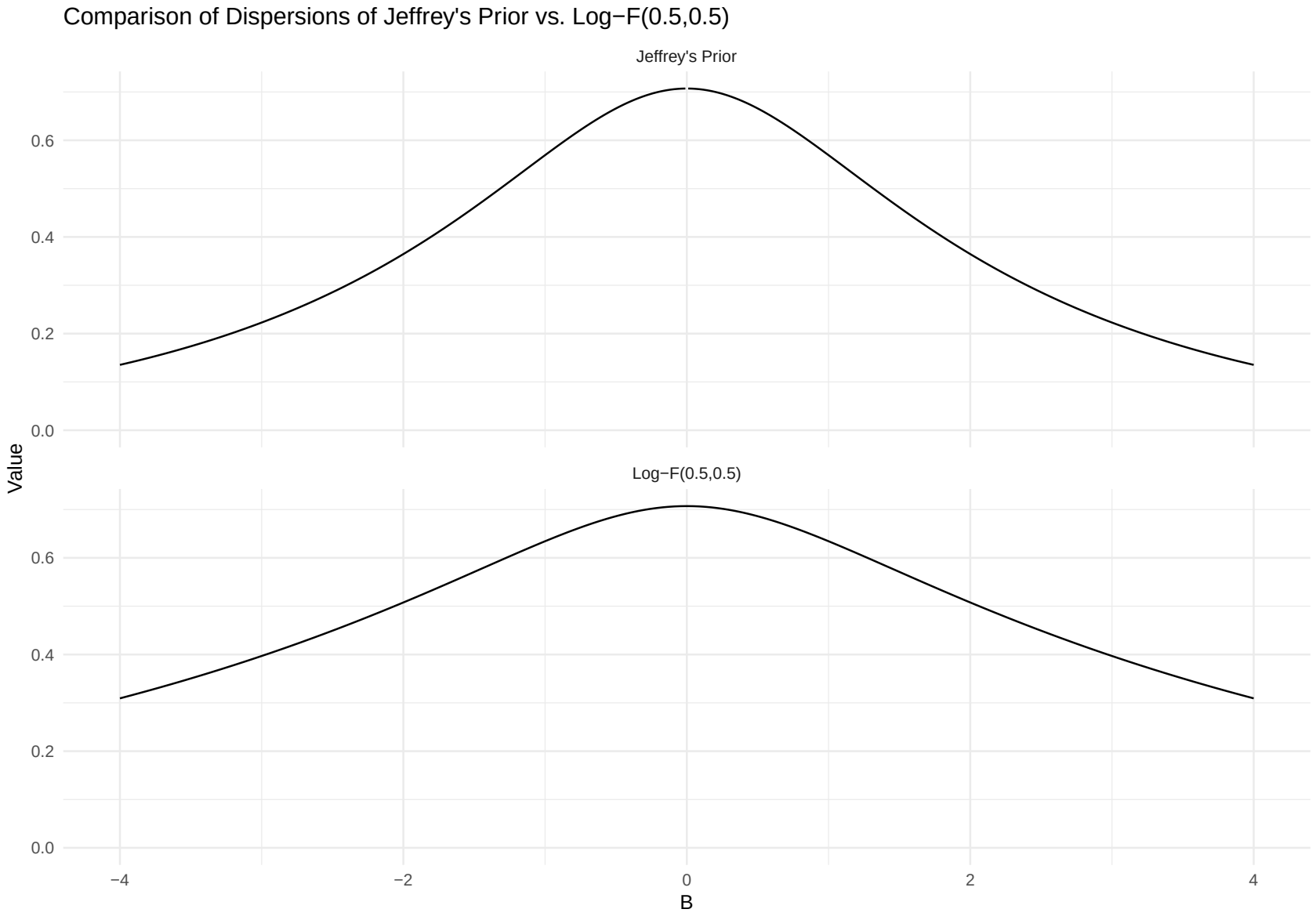
$$L^*(\beta_0) \propto \; L(\beta_0) * \Big[\frac{n\, e^{\beta_0}}{\left(1 + e^{\beta_0}\right)^2}\Big]^{1/2}$$

We consider a more adaptive prior instead:

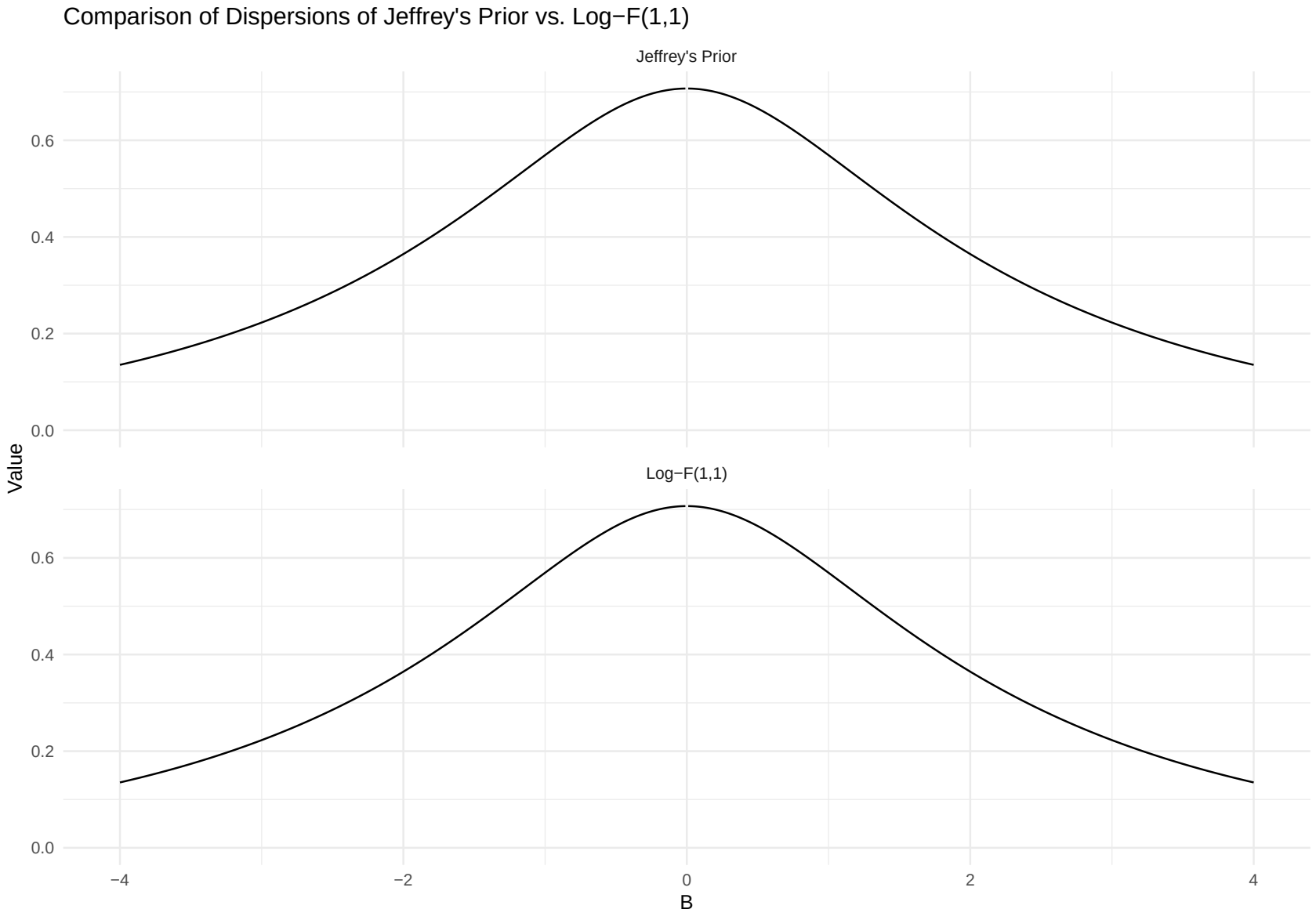
$$L^*(\beta_0) \propto \; L(\beta_0) * \Big[\frac{e^{\beta_0}}{\left(1 + e^{\beta_0}\right)^2}\Big]^{m/2}$$

Called the log-F(m, m) prior.

Adjustable Dispersion



Adjustable Dispersion



Adjustable Dispersion

Larger m pulls $\hat{\beta}_0$ towards 0 more strongly.

Pseudo-Data

This penalized likelihood

$$L(\beta_0) \propto \; L(\beta_0) * F(\beta_0)^{m/2}$$

has a closed-form maximum at

$$\hat{\beta}_0 \; = \; \log\Big(\frac{y+m/2}{n-y+m/2}\Big).$$

which is mathematically equivalent to adding m/2 successes and m/2 failures to your regular data, and obtaining the MLE.

Multi-Parameter Logistic Regression

For a multiparameter logistic model:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^{\mathsf{T}}\,\boldsymbol{\beta},$$

Firth's Adjusted Likelihood is:

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) * |F(\boldsymbol{\beta})|^{1/2}$$

Where Fisher's Information is now a function of the design matrix, and is centred around 0.

Argument for Selective Penalization

Arguments against Jeffrey's Prior:

- Jeffrey's Prior may not be ideal when true coefficients are far from 0, as it tends to overshrink the estimates.
- It is not a true prior in the sense that it does not make use of prior knowledge.
- Penalizing the intercept term is also not recommended, to avoid disrupting baseline-odds.

Arguments for independent log-F(m, m) priors:

- We can adjust the strength of penalization m for each coefficient based on prior knowledge
- We can avoid penalizing the intercept coefficient

Log-F Prior

$$L^*(\beta) \propto \; L(\beta) * \Big[\frac{e^{\beta_j}}{\left(1+e^{\beta_j}\right)^2}\Big]^{m_j/2}$$

If we wish to penalize β_j of β, we can add m/2 successes and m/2 failures, where the corresponding $x_j = 1$ and all other covariates $=$ 0, *including the intercept constant*.

Implementation Example

Example data:

```
##   y          X1         X2
## 1 1  3.1905643  0.7488708
## 2 0 -0.3913190  3.9831363
## 3 1  2.6260029  3.4764541
## 4 1  2.0850844 -0.2705859
## 5 0 -0.2196981 -3.6611475
## 6 0 -1.7602619  1.7459681
```

Adding Pseudo Data

Example for applying a m=3 penalty to β_1

```
#add a manual intercept that we can set to 0
df$intercept = 1

#we will fit the glm using the weight method, assign a weight of 1 to existing data
df$weight = 1

m=3

#add_row function is from tibble/tidyverse
df <- add_row(df,
              y=0.5,   #y represents the proportion of successes out of weight trials
              X1 = 1,
              X2 = 0,
              intercept = 0,
              weight = m)

model <- glm(y ~ -1 + intercept + X1 + X2,
              family=binomial,
              data=df,
              weight = weight)
```

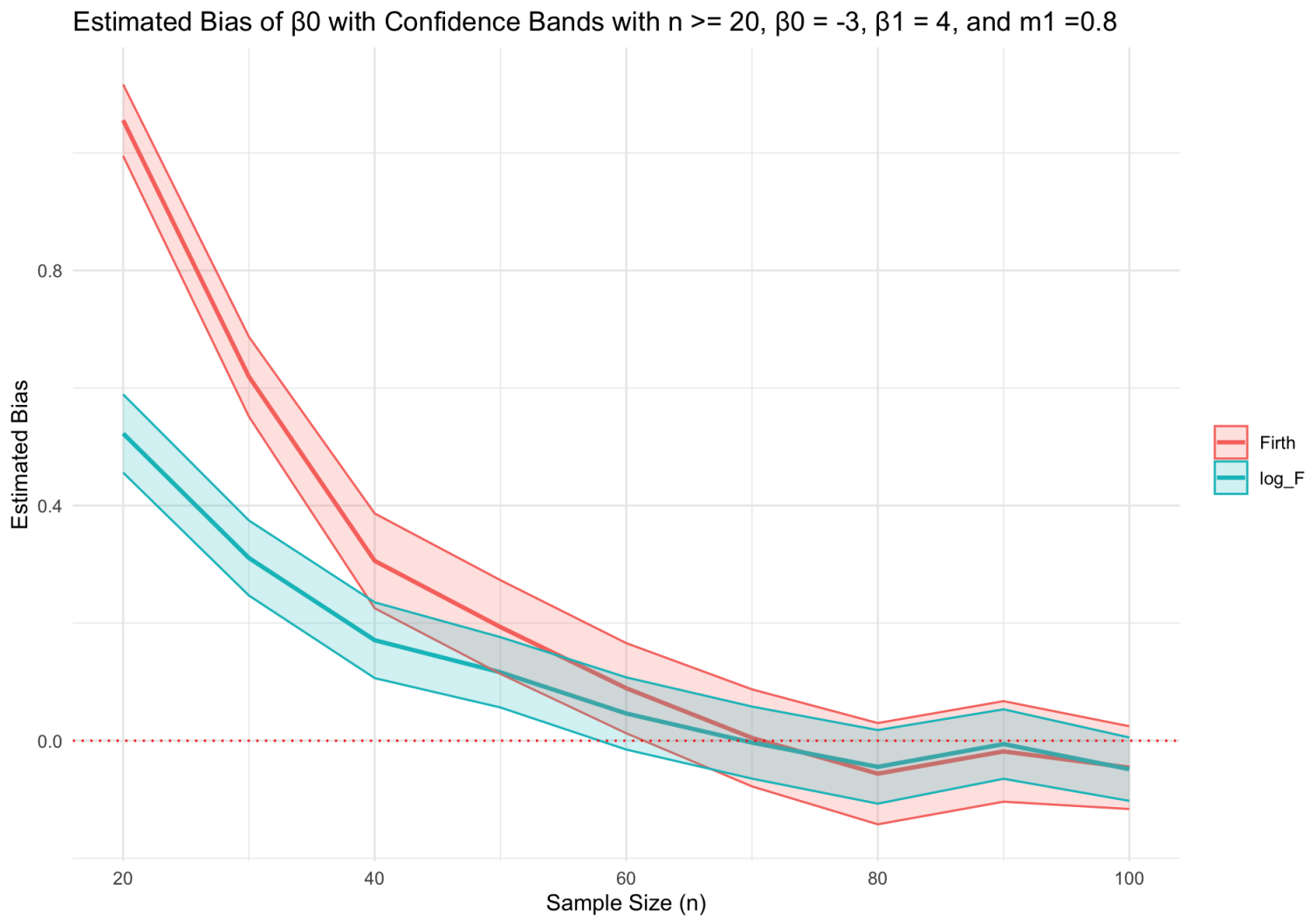
```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

How to Pick m?

- Based on prior beliefs about e^{β_j} .
- If we believe a one-unit increase in x_j multiples the log odds by a factor of roughly at most 2, we can set this as an upper-bound of a 95% confidence interval
- Find a F(m, m) distribution such that $P(2 > F) \approx 0.025$

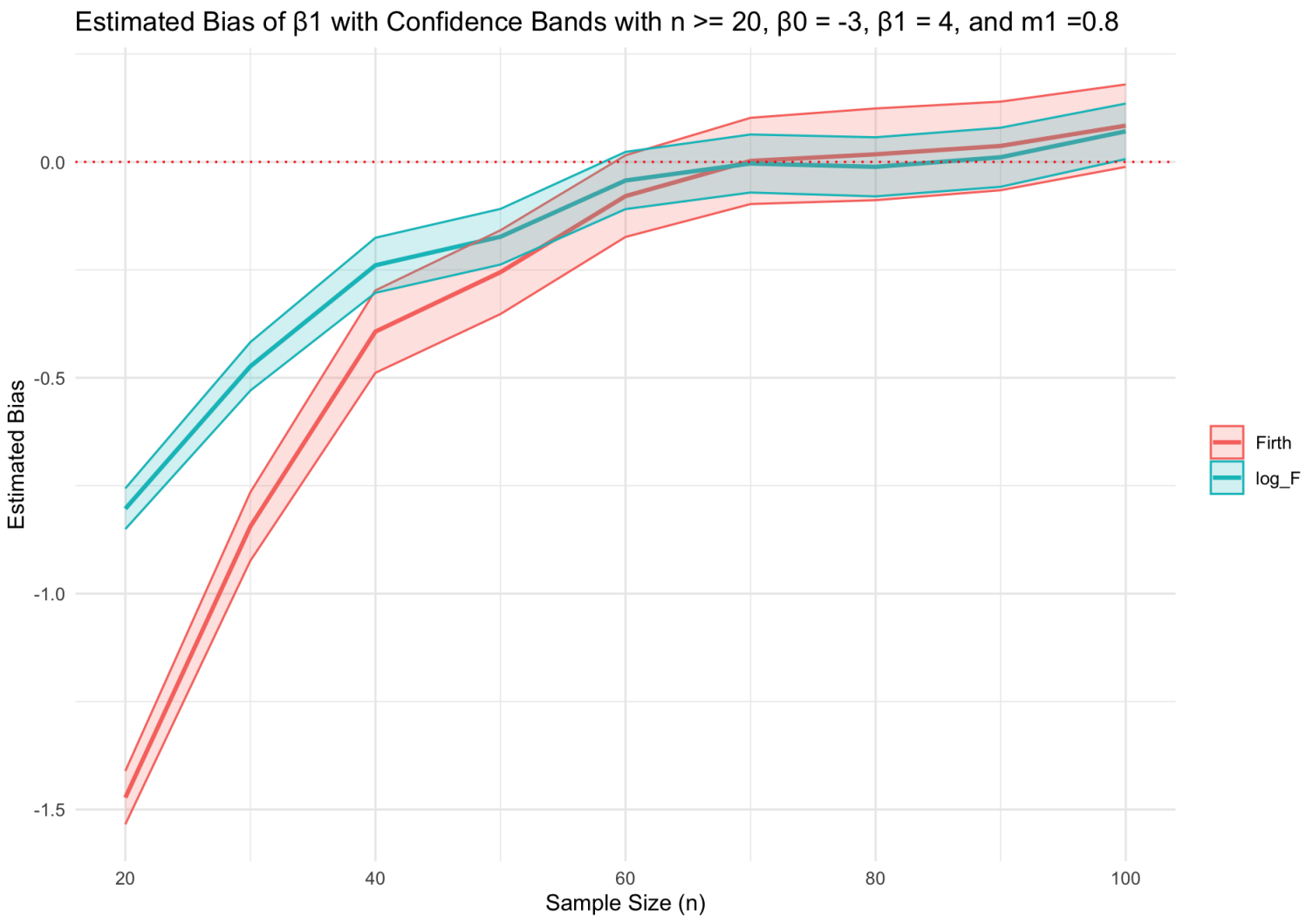
Larger values of β will result in smaller m!

Influence of m



We set $m_1 = 0.8$ by assuming β_1 is, at most, ~ 8

Influence of m

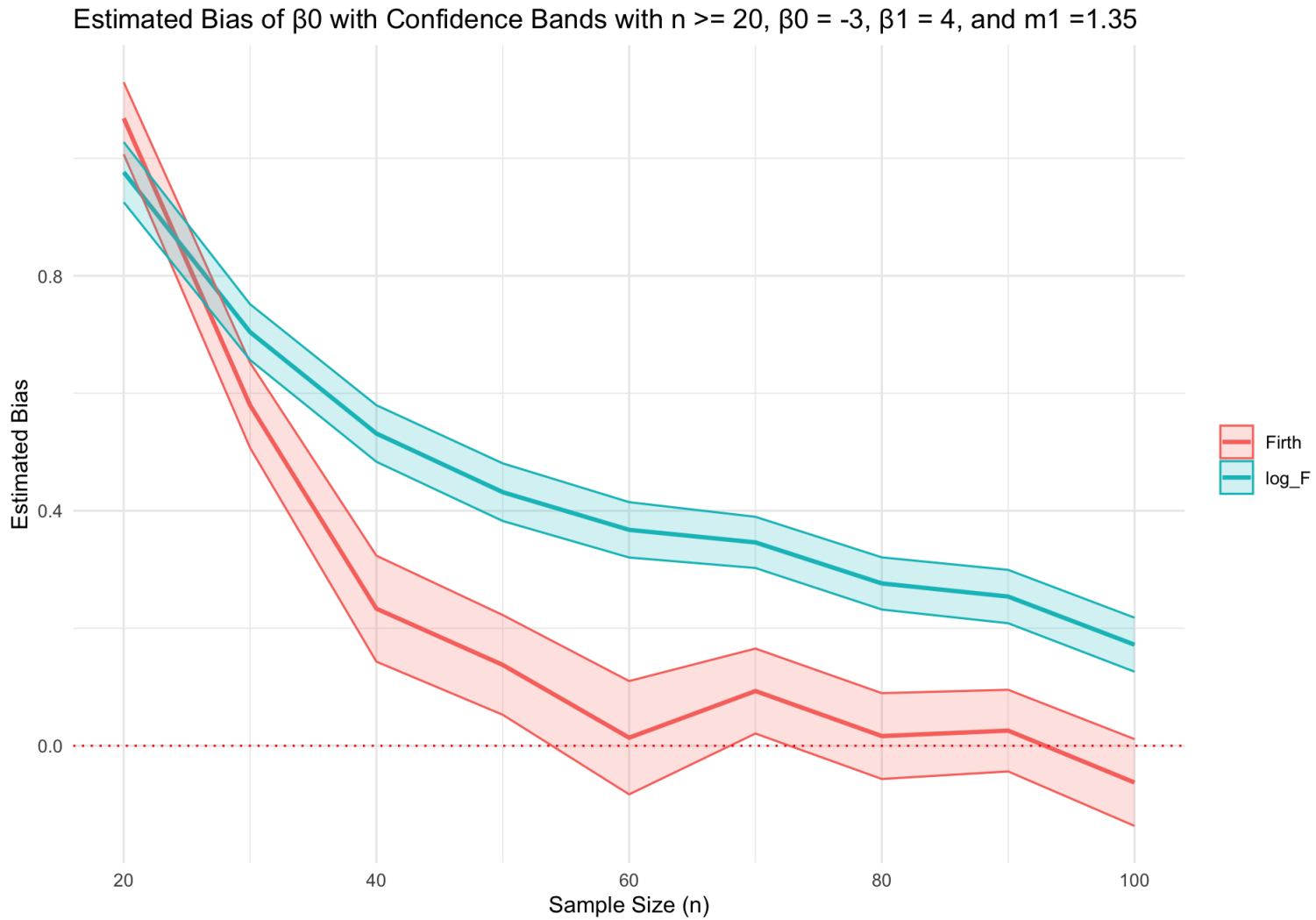


We set $m_1 = 0.8$ by assuming β_1 is, at most, ~ 8

Influence of m

What if we mistakenly believe that β_1 is relatively small, and rarely exceeds 5?

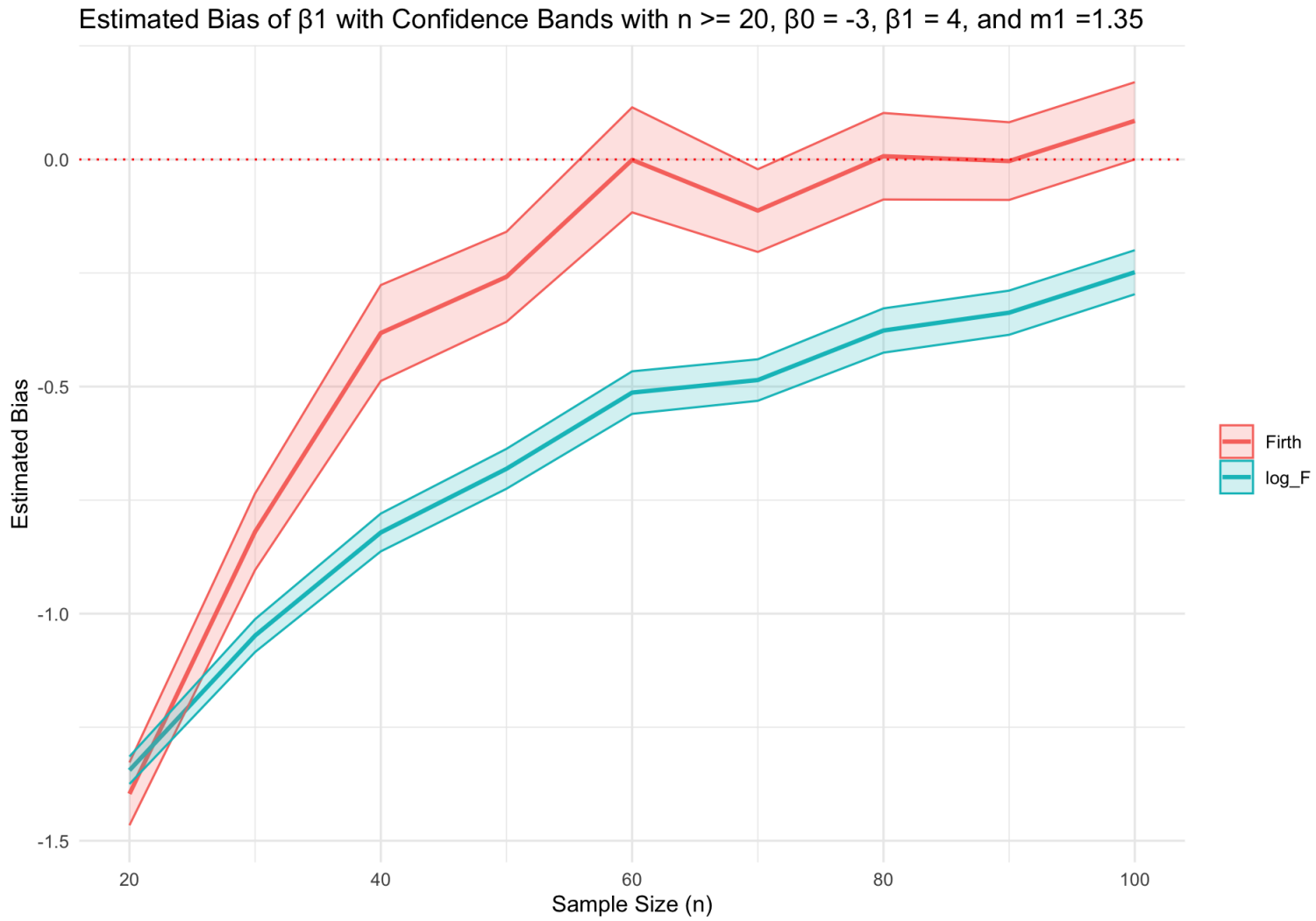
Our estimates are overshrunk!



Influence of m

What if we mistakenly believe that β_1 is relatively small, and rarely exceeds 5?

Our estimates are overshrunk!



Influence of m

Incorrect prior knowledge can add additional bias to your estimates.

If no prior knowledge is known, Greenland & Mansournia proposes a m=1 penalty on all coefficients except β_0.

This is the default we will use in comparisons of all 3 bias-reduction methods.

Simulation Study

Simulation Study

Study Setting

- **Model**: $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{1i}$.
- $X_{1i} \sim N(0, 1)$.
- All Confidence Intervals are Wald CIs.
- **Terminology**:
  1. **Easy Case**: Most observations have π_i near 0.5.
  2. **Moderate Case**: Most observations have π_i around 0.2 to 0.4 or 0.6 to 0.8.
  3. **Extreme Case**: Most observations have π_i around 0.01 to 0.1 or 0.9 to 0.999.
- Each simulation result is based on 1000 datasets generated from same true model.

Simulation Study

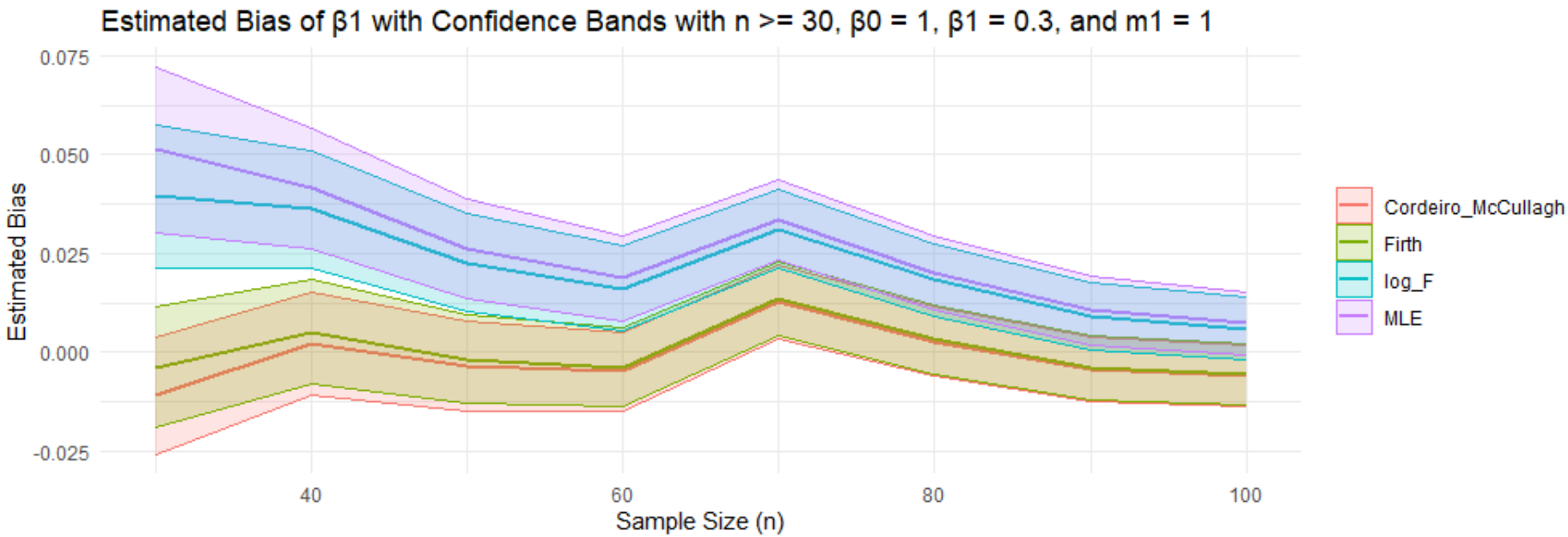


Figure: Easier case

Simulation Study

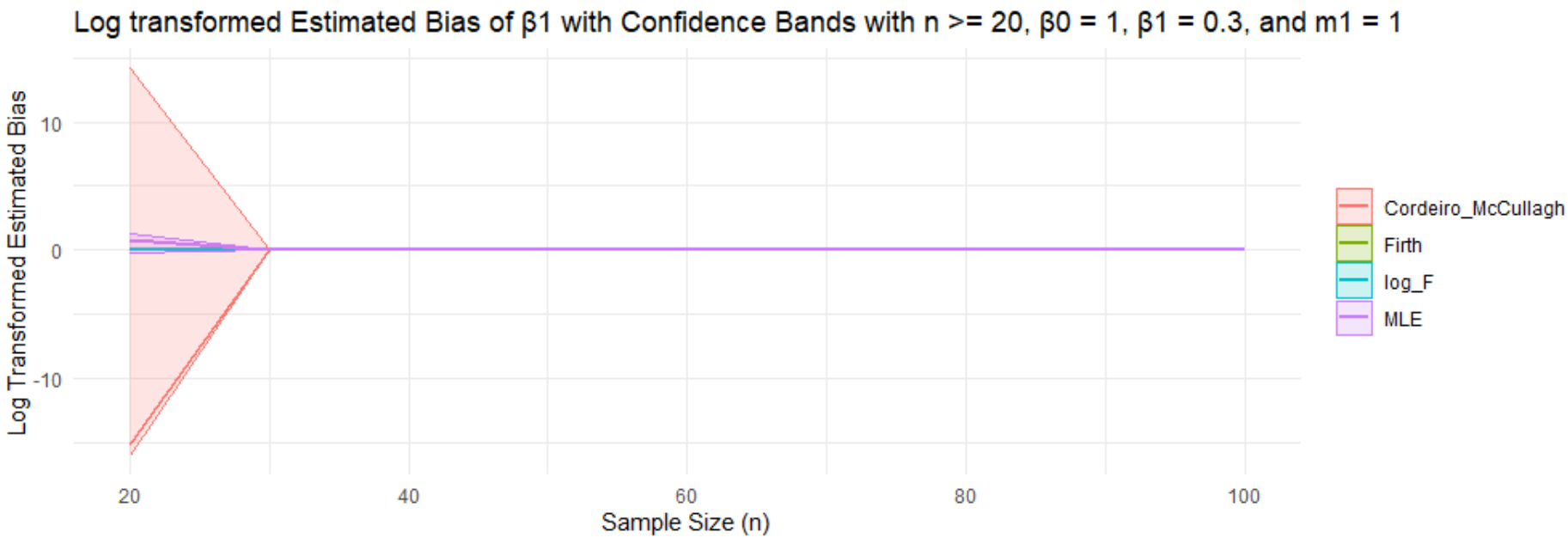


Figure: Easier case when sample size is less than 30

Simulation Study

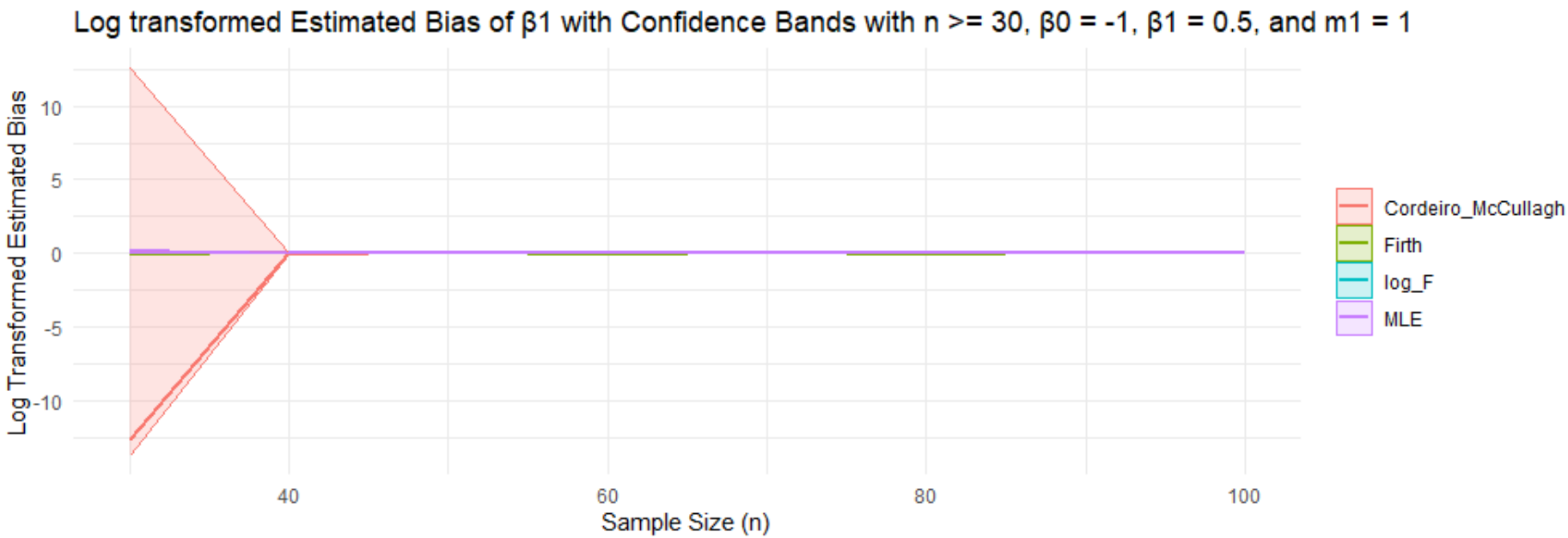


Figure: Moderate case

Simulation Study

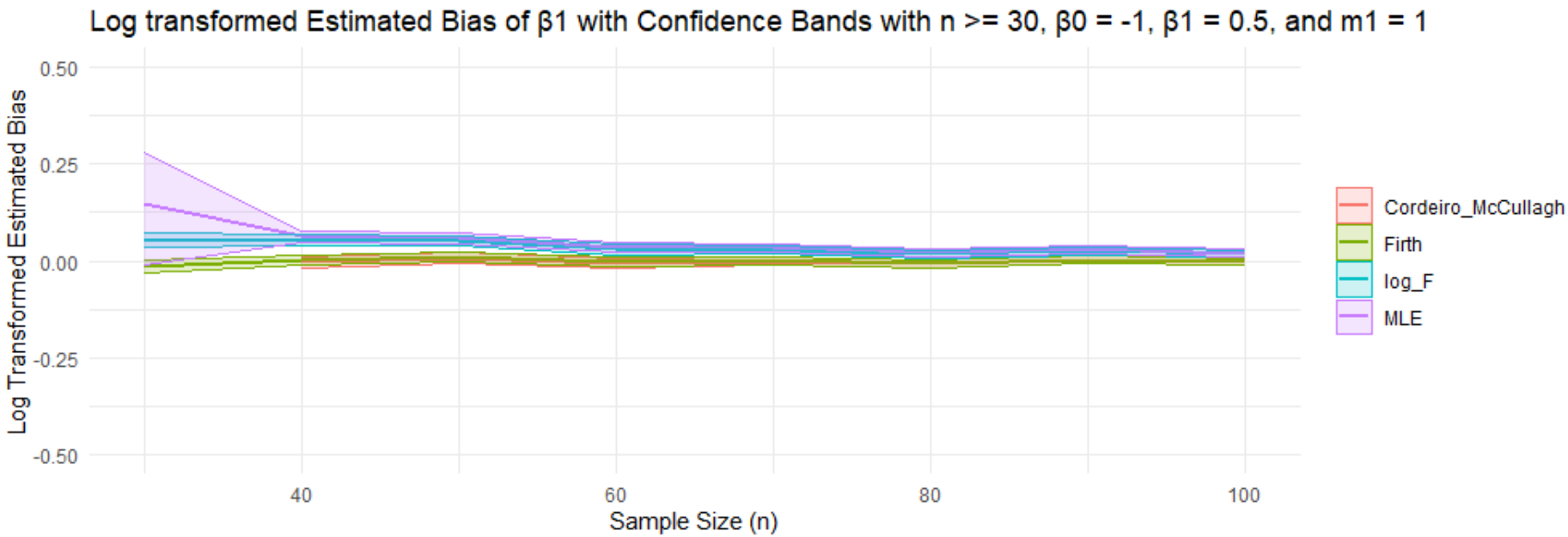


Figure: Moderate case - a closer look

Simulation Study

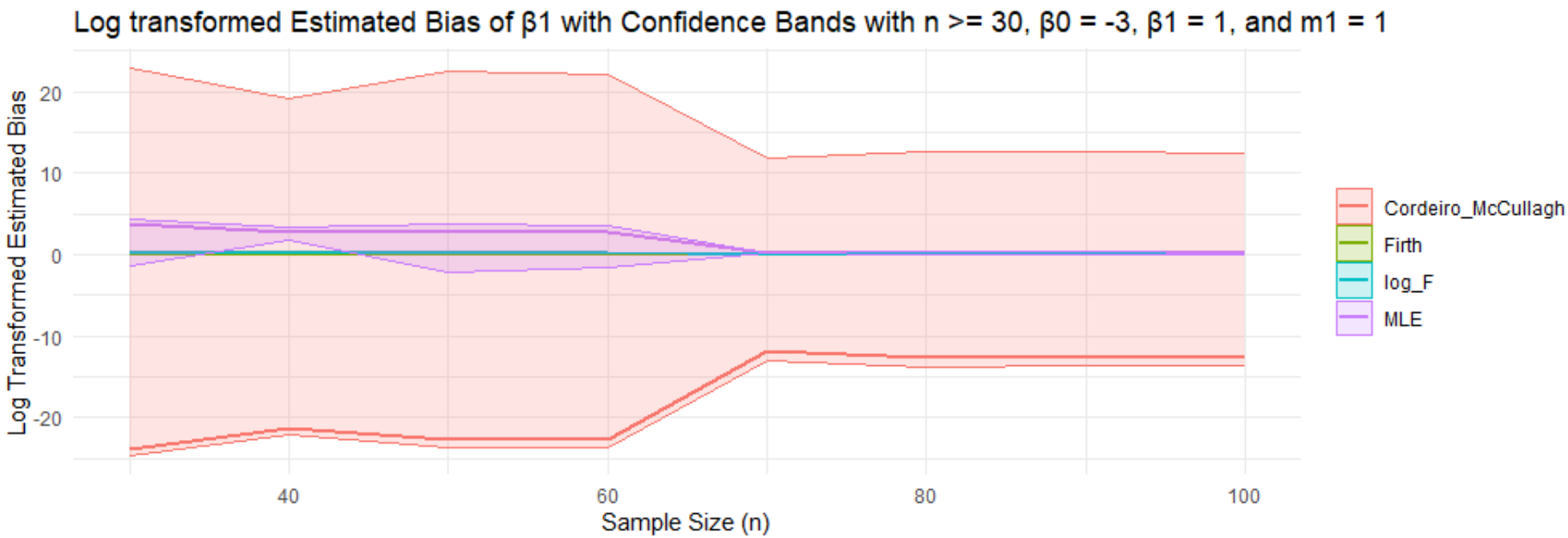


Figure: Extreme Case

Simulation Study

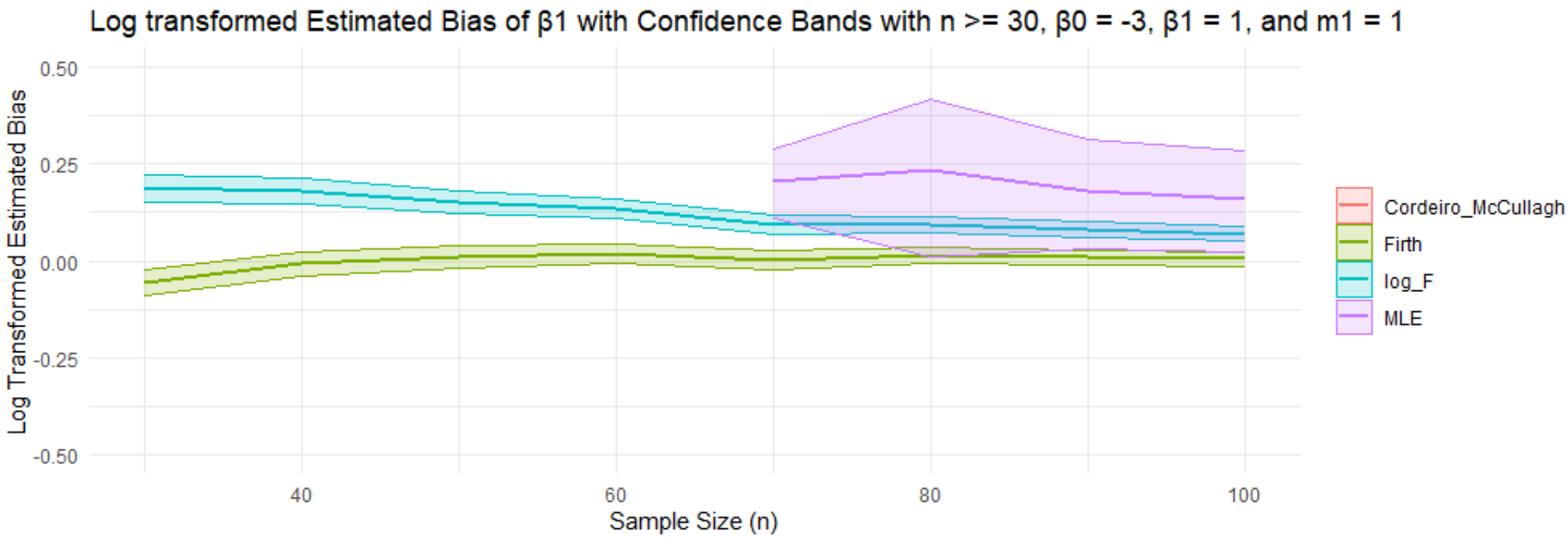


Figure: Extreme Case - A closer look

Simulation Study

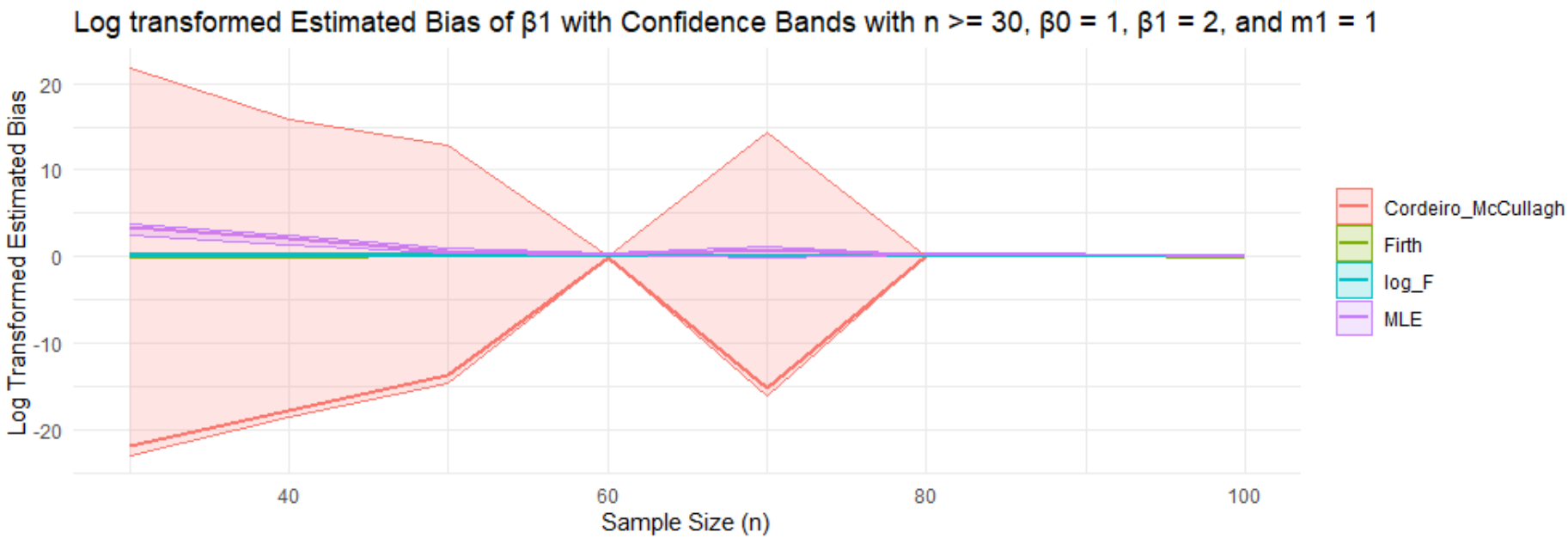


Figure: An interesting case

Simulation Study

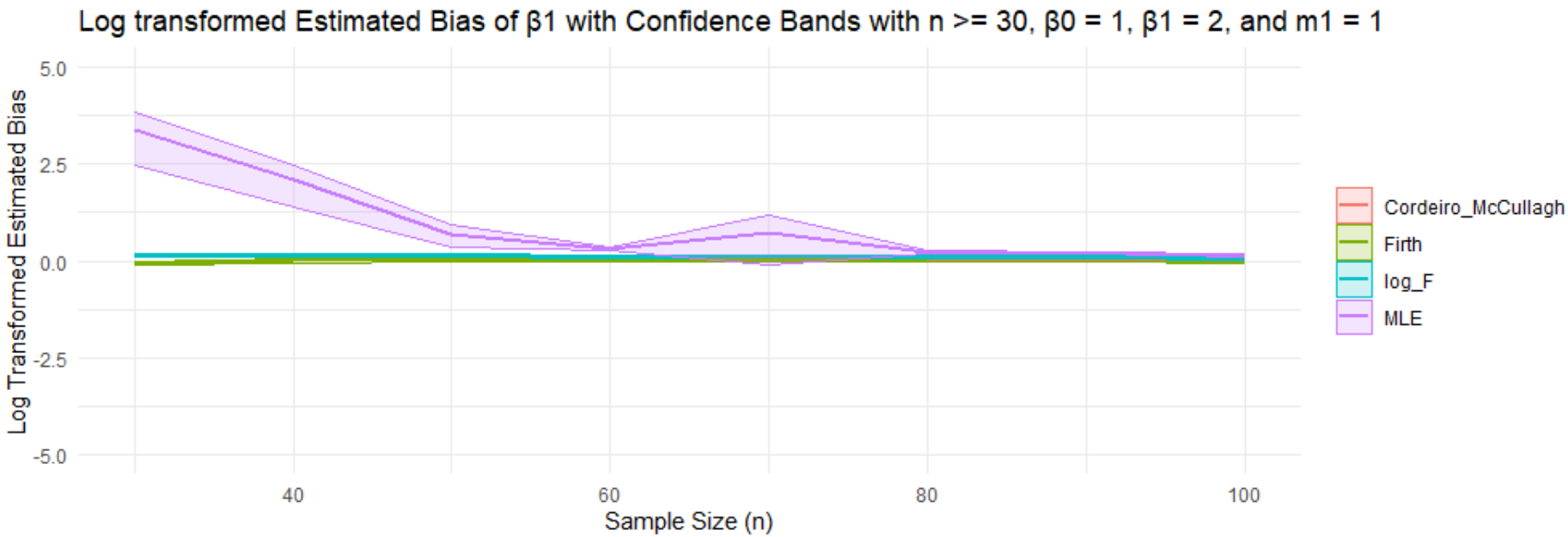


Figure: An Interesting Case - A closer look

Simulation Study

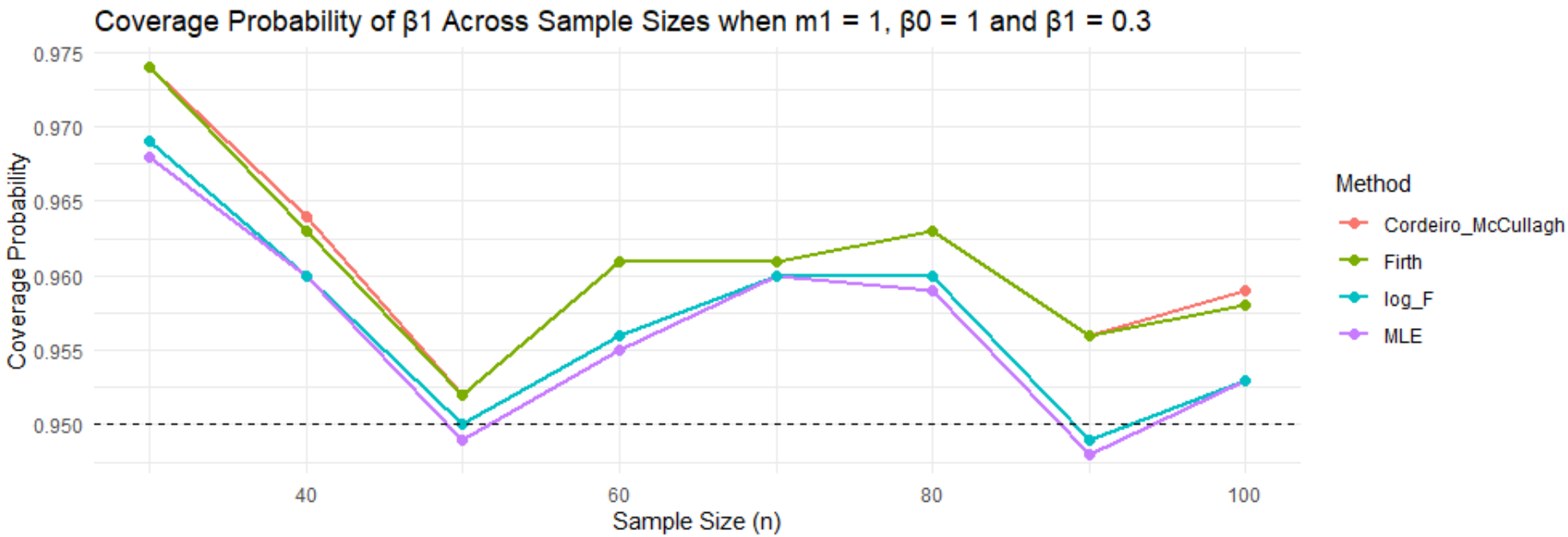


Figure: Easier Case

Simulation Study

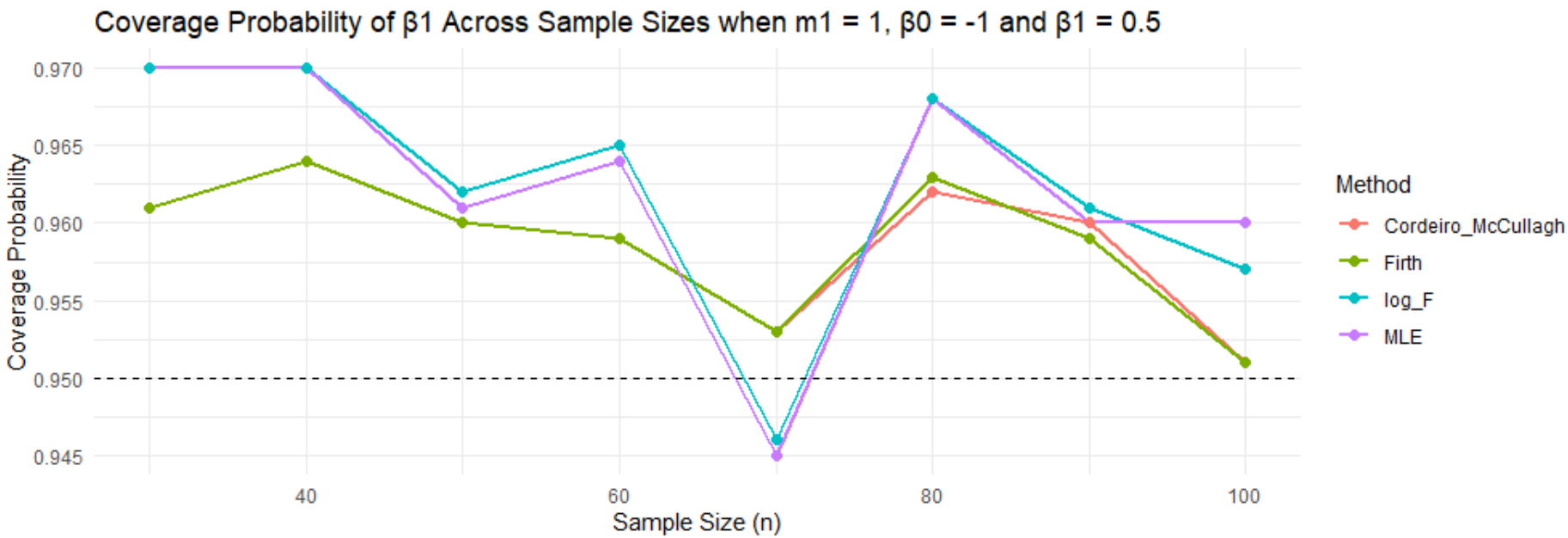


Figure: Moderate Case

Simulation Study

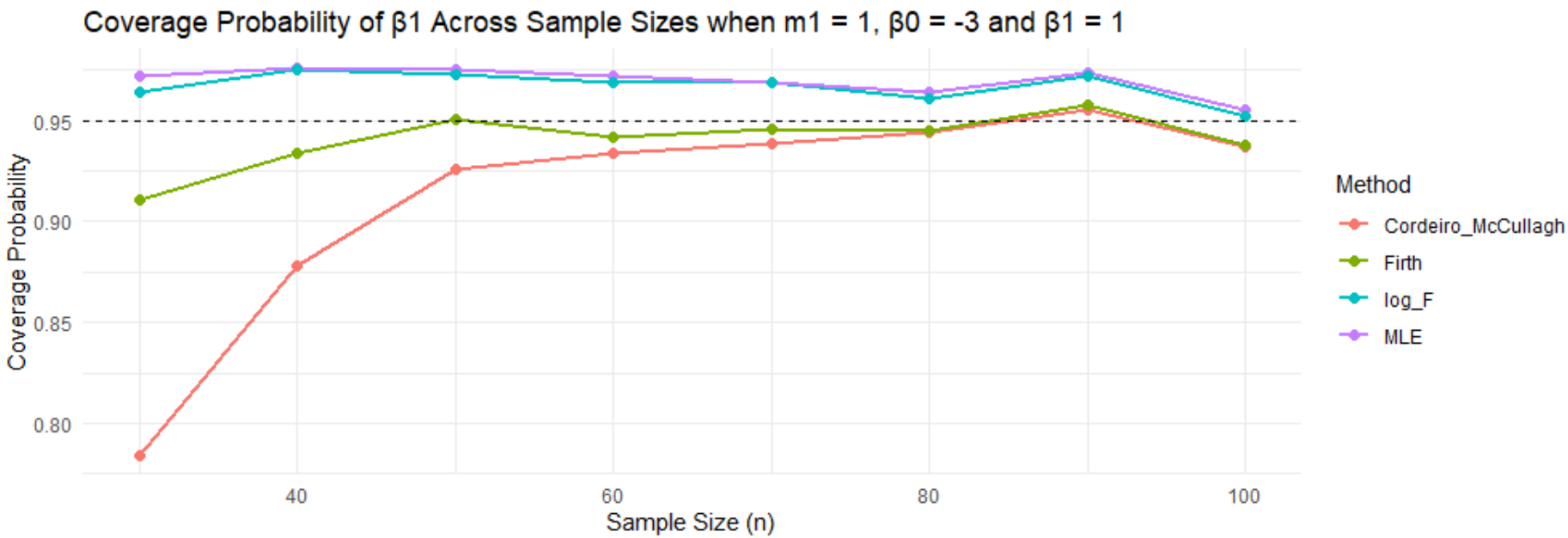


Figure: Extreme Case

Key Takeaways

- Asymptotic Bias Correction seems to have smaller bias only when most of the true probabilities are near 0.5 and sample size is ≥ 30. It performs worse than the MLE in moderate and extreme cases with small sample sizes.
- Penalized Likelihood methods are more stable for both smaller sample sizes and extreme β's.
- Firth's approach tend to perform better than the default $\log F(1,1)$ applied to all coefficients except the intercept when no prior knowledge is available.
- $\log -F$ prior may be preferable when sample sizes are low and strong prior knowledge about β is known.

Thank You!