

Ch. 1 - Model comparisons

Wade VanderWright

2022-02-25

Contents

| | |
|--|----------|
| Model Comparisons | 1 |
| Model 1 - CFAR ~ GSH | 1 |
| Model 2 - CFAR ~ GSH * Max Size | 4 |
| Model 3 - CFAR ~ GSH + Max Size (No interaction) | 6 |
| Model 4 - CFAR ~ GSH + (1 Max Size) (No interaction) | 9 |
| Model 5 - CFAR ~ GSH + Phylogeny | 11 |
| Model 5 - CFAR ~ GSH * Max Size + Phylogeny | 15 |
| Model 6 - CFAR ~ GSH + Max Size + Phylogeny | 20 |

Model Comparisons

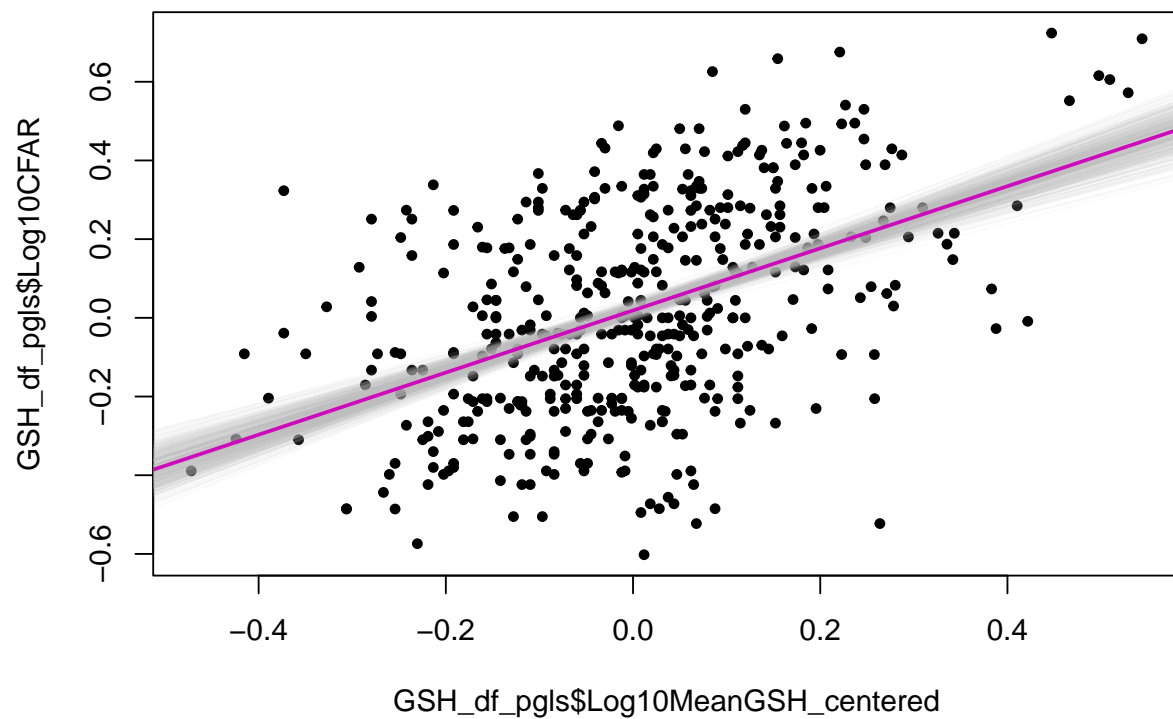
Model 1 - CFAR ~ GSH

brms model

```
Model_plain <- brm(Log10CFAR ~ Log10MeanGSH_centered,  
  data = GSH_df_pgls,  
  family = gaussian(),  
  prior = Plain_prior  
)
```

```
fixef(Model_plain, summary = T)
```

```
##              Estimate Est.Error      Q2.5      Q97.5  
## Intercept          0.01878724 0.01121401 -0.003048758 0.04068312  
## Log10MeanGSH_centered 0.78943177 0.06909079  0.653733169 0.92180824
```



STAN model

Data considered

```
data {
  int <lower=1> N;
  vector[N] x;
  vector[N] y;
}
```

The parameters accepted by the model

```
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;
}
```

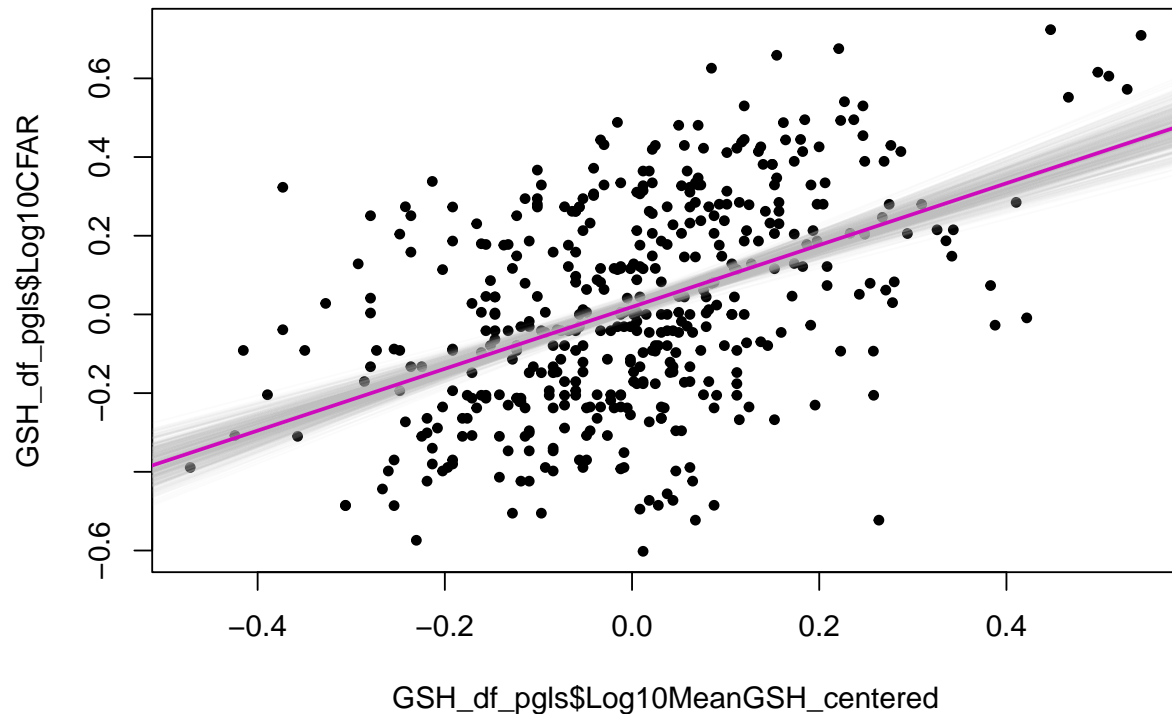
Model form:

```
model {
  sigma ~ student_t(3, 0, 10);
```

```
y ~ normal(alpha + x * beta , sigma);
}
```

```
fit1
```

```
## Inference for Stan model: StanModel1.
## 4 chains, each with iter=1000; warmup=500; thin=1;
## post-warmup draws per chain=500, total post-warmup draws=2000.
##
##          mean se_mean   sd  2.5%  25%   50%   75%  97.5% n_eff Rhat
## alpha    0.02    0.00 0.01   0.00  0.01  0.02  0.03   0.04 2478   1
## beta     0.79    0.00 0.07   0.65  0.74  0.79  0.83   0.92 1573   1
## sigma    0.24    0.00 0.01   0.22  0.23  0.24  0.24   0.25 2327   1
## lp__    429.85    0.04 1.24 426.46 429.30 430.13 430.76 431.26 1072   1
##
## Samples were drawn using NUTS(diag_e) at Thu Mar 03 10:00:20 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```



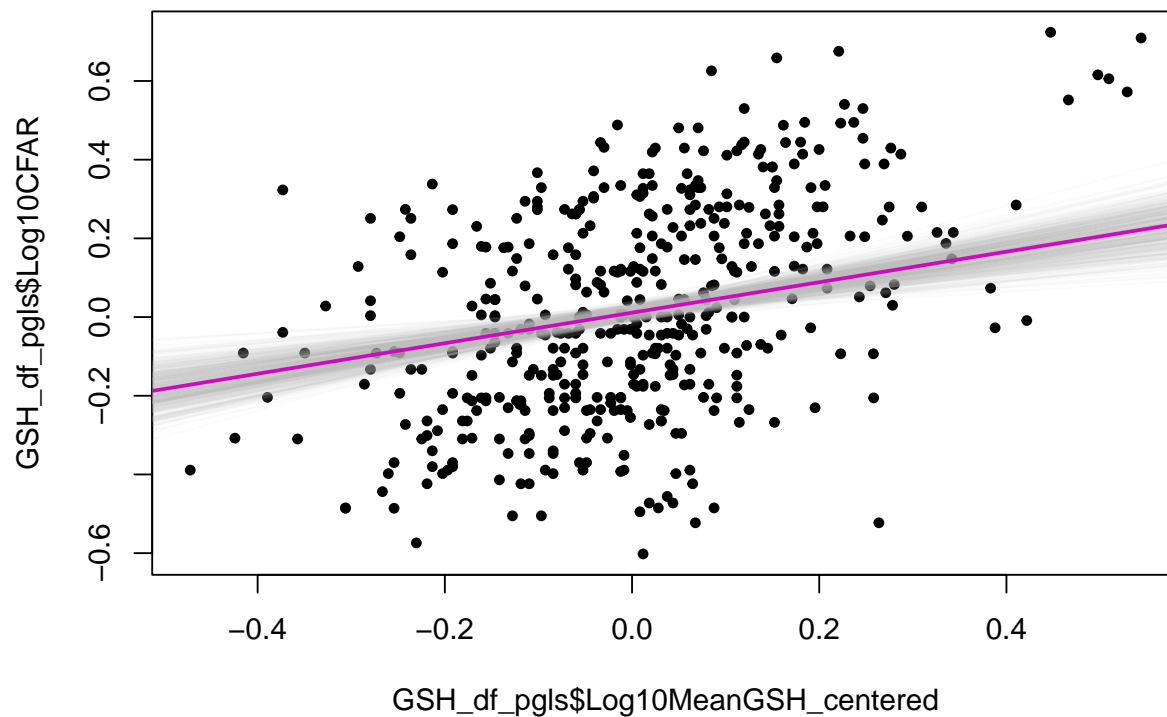
Model 2 - CFAR ~ GSH * Max Size

brms model

```
Model_size <- brm(Log10CFAR ~ Log10MeanGSH_centered * Log10MaxSize_centered,  
  data = GSH_df_pgls,  
  family = gaussian(),  
  prior = Size_prior  
)
```

```
fixef(Model_size, summary = T)
```

| ## | Estimate | Est.Error | Q2.5 |
|--|------------|------------|-------------|
| ## Intercept | 0.01107801 | 0.01158262 | -0.01069365 |
| ## Log10MeanGSH_centered | 0.38761087 | 0.08262037 | 0.22932179 |
| ## Log10MaxSize_centered | 0.30460627 | 0.04281846 | 0.22222482 |
| ## Log10MeanGSH_centered:Log10MaxSize_centered | 0.26628115 | 0.16406086 | -0.05284798 |
| ## | Q97.5 | | |
| ## Intercept | 0.03436049 | | |
| ## Log10MeanGSH_centered | 0.55212619 | | |
| ## Log10MaxSize_centered | 0.38716498 | | |
| ## Log10MeanGSH_centered:Log10MaxSize_centered | 0.58368860 | | |



STAN model

Data considered

```
data {  
  int <lower=0> N; // number of data points  
  int <lower=0> K; // number of predictors  
  matrix[N,K] x; // predictor matrix  
  vector[N] y; // CFAR  
}
```

The parameters accepted by the model.

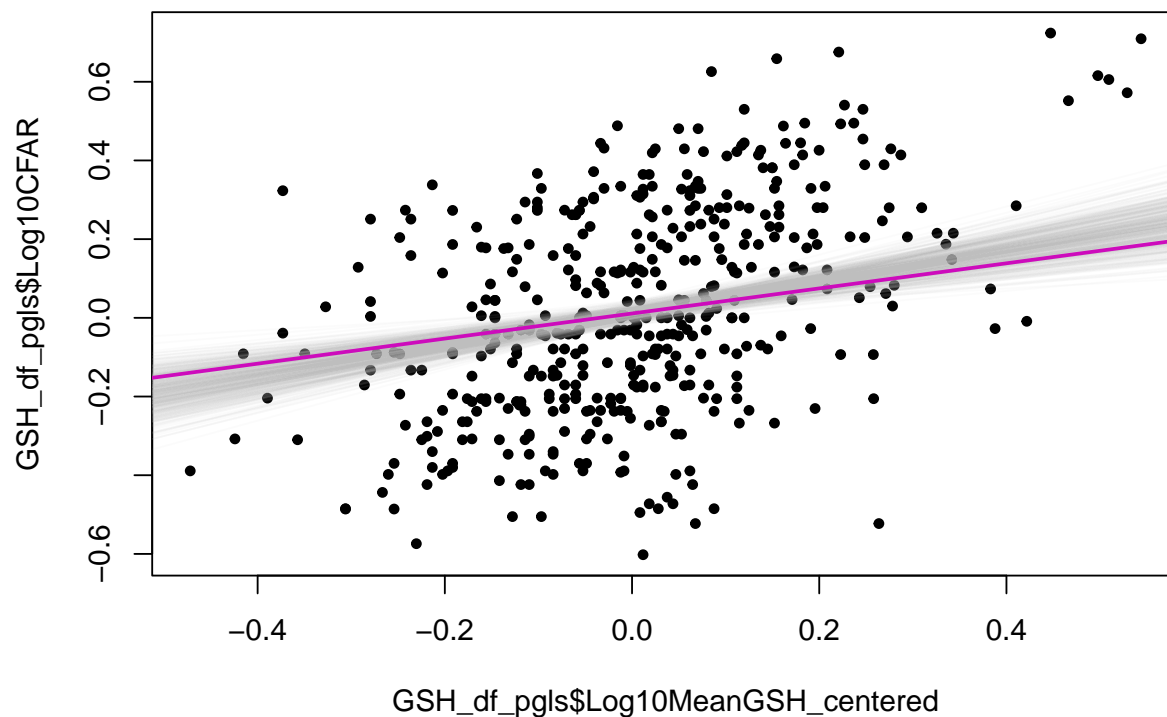
```
parameters {  
  real alpha;  
  vector[K] beta;  
  real<lower=0> sigma;  
}
```

The model to be estimated. We model the output ‘y’ to be normally distributed with mean ‘mu’ and standard deviation ‘sigma’.

```
model {  
  
  sigma ~ student_t(3, 0, 10);  
  
  y ~ normal(alpha + x * beta , sigma);  
}
```

fit2

```
## Inference for Stan model: StanModel2.  
## 4 chains, each with iter=1000; warmup=500; thin=1;  
## post-warmup draws per chain=500, total post-warmup draws=2000.  
##  
##           mean se_mean   sd  2.5%  25%   50%   75%  97.5% n_eff Rhat  
## alpha      0.01    0.00 0.01  -0.01   0.00   0.01   0.02   0.04  1841   1  
## beta[1]    0.39    0.00 0.08   0.23   0.33   0.39   0.44   0.55  1758   1  
## beta[2]    0.31    0.00 0.04   0.22   0.28   0.31   0.33   0.39  1742   1  
## beta[3]    0.26    0.00 0.16  -0.06   0.16   0.27   0.37   0.59  1975   1  
## sigma      0.22    0.00 0.01   0.21   0.22   0.22   0.23   0.24  2516   1  
## lp__      459.42    0.05 1.62 455.40 458.60 459.75 460.60 461.57   884   1  
##  
## Samples were drawn using NUTS(diag_e) at Wed Mar 02 11:20:54 2022.  
## For each parameter, n_eff is a crude measure of effective sample size,  
## and Rhat is the potential scale reduction factor on split chains (at  
## convergence, Rhat=1).
```



Model 3 - CFAR \sim GSH + Max Size (No interaction)

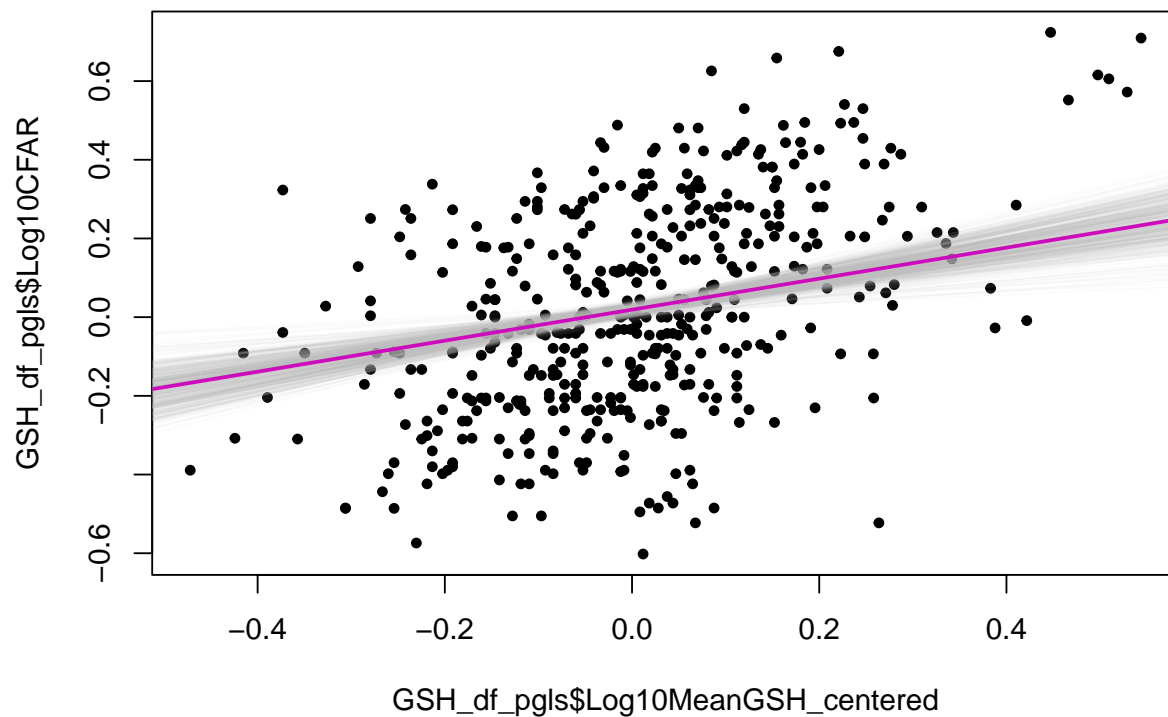
brms model

```
Model_size_2 <- brm(Log10CFAR ~ Log10MeanGSH_centered + Log10MaxSize_centered,
  data = GSH_df_pgls,
  family = gaussian(),
  prior = Size_prior_2)
```

```
summary(Model_size_2)
```

```
## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: Log10CFAR ~ Log10MeanGSH_centered + Log10MaxSize_centered
## Data: GSH_df_pgls (Number of observations: 456)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
## total post-warmup draws = 4000
##
## Population-Level Effects:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS
## Intercept           0.02     0.01  -0.00   0.04 1.00    3526
## Log10MeanGSH_centered 0.39     0.08   0.24   0.55 1.00    3567
## Log10MaxSize_centered 0.33     0.04   0.25   0.40 1.00    3528
```

```
##                               Tail_ESS
## Intercept                    2556
## Log10MeanGSH_centered       2768
## Log10MaxSize_centered       3043
##
## Family Specific Parameters:
##      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      0.22      0.01   0.21   0.24 1.00   3652   2981
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```



STAN model

Data considered

```
data {
  int <lower=0> N; // number of data points
  int <lower=0> K; // number of predictors
  matrix[N,K] x; // predictor matrix
  vector[N] y; // CFAR
}
```

The parameters accepted by the model.

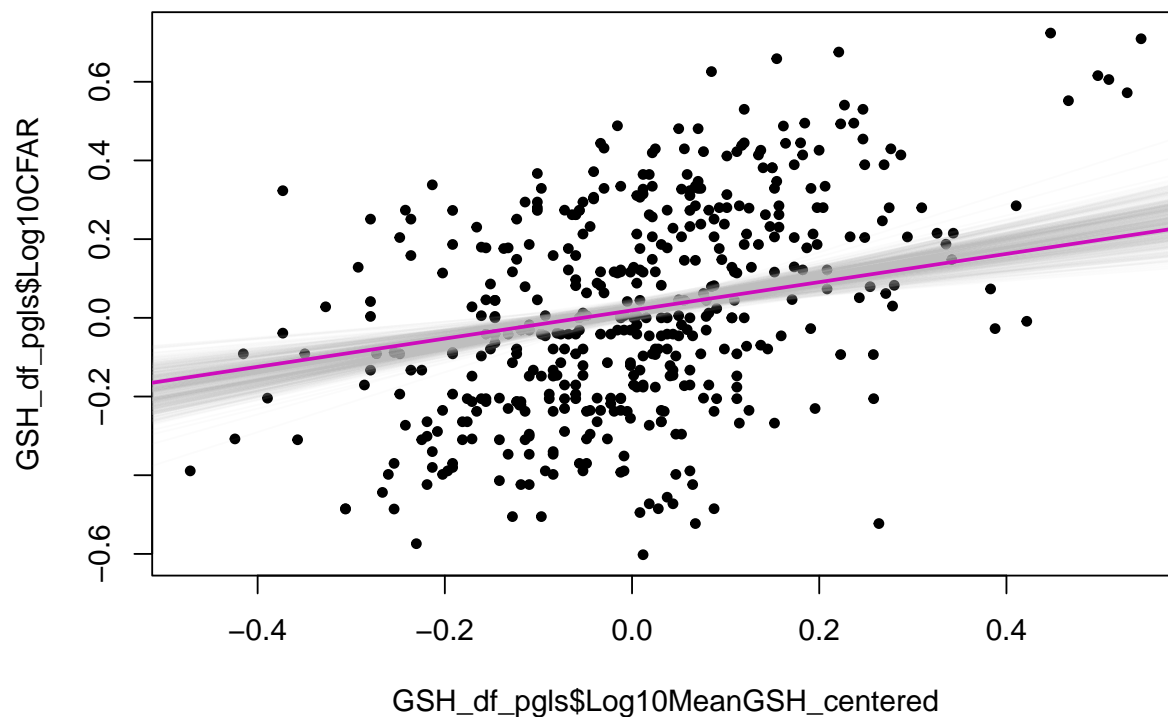
```
parameters {  
  real alpha;  
  vector[K] beta;  
  real<lower=0> sigma;  
}
```

The model to be estimated. We model the output ‘y’ to be normally distributed with mean ‘mu’ and standard deviation ‘sigma’.

```
model {  
  
  sigma ~ student_t(3, 0, 10);  
  
  y ~ normal(alpha + x * beta , sigma);  
}
```

```
fit3
```

```
## Inference for Stan model: StanModel2.  
## 4 chains, each with iter=1000; warmup=500; thin=1;  
## post-warmup draws per chain=500, total post-warmup draws=2000.  
##  
##          mean se_mean   sd  2.5%   25%   50%   75%  97.5% n_eff Rhat  
## alpha      0.02    0.00 0.01   0.00   0.01   0.02   0.03   0.04  2075    1  
## beta[1]     0.39    0.00 0.08   0.23   0.34   0.39   0.45   0.55  1620    1  
## beta[2]     0.32    0.00 0.04   0.24   0.30   0.32   0.35   0.40  1550    1  
## sigma       0.22    0.00 0.01   0.21   0.22   0.22   0.23   0.24  1790    1  
## lp__       458.67    0.05 1.41 455.06 458.03 458.99 459.70 460.42   948    1  
##  
## Samples were drawn using NUTS(diag_e) at Wed Mar 02 11:22:15 2022.  
## For each parameter, n_eff is a crude measure of effective sample size,  
## and Rhat is the potential scale reduction factor on split chains (at  
## convergence, Rhat=1).
```

Model 4 - CFAR \sim GSH + (1|Max Size) (No interaction)

brms model

```
Model_size_3 <- brm(Log10CFAR ~ Log10MeanGSH_centered + (1|Log10MaxSize_centered),
  data = GSH_df_pgls,
  family = gaussian(),
  prior = Size_prior_3,
  control = list(adapt_delta = 0.9)) #needed to make model converge ...
'''

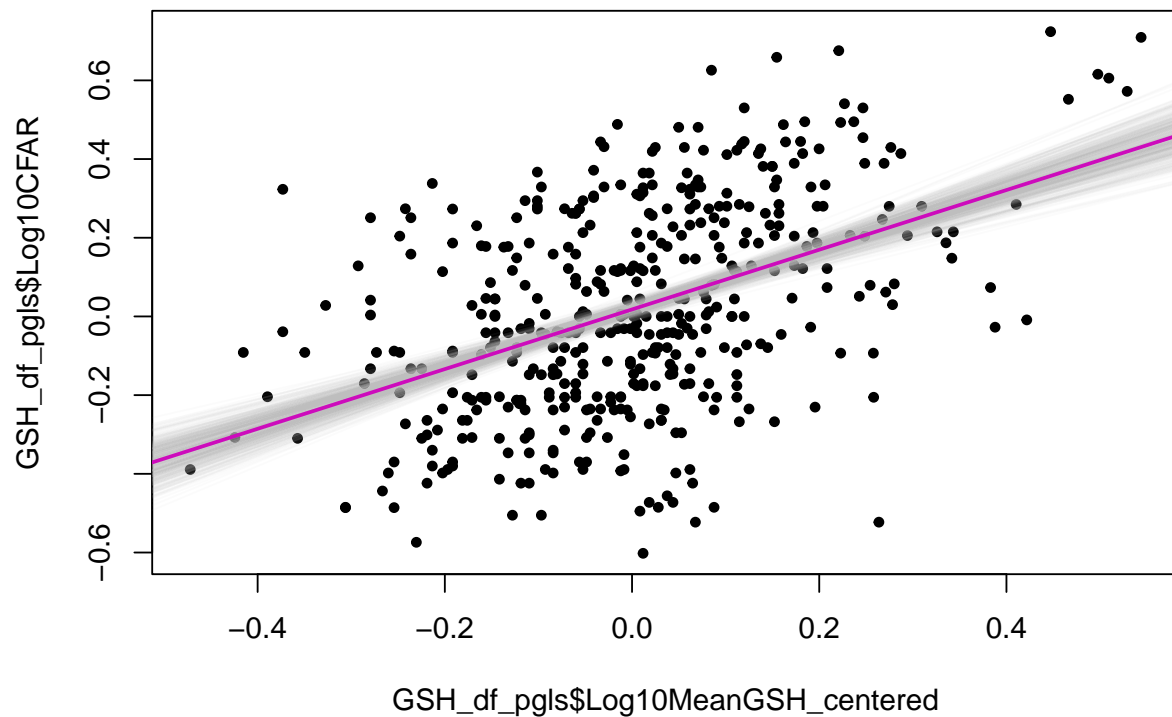
'''r
summary(Model_size_3)

## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: Log10CFAR ~ Log10MeanGSH_centered + (1 | Log10MaxSize_centered)
## Data: GSH_df_pgls (Number of observations: 456)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
## total post-warmup draws = 4000
##
## Group-Level Effects:
## ~Log10MaxSize_centered (Number of levels: 222)
```

```

##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)    0.06     0.03    0.01    0.11 1.01     435     525
##
## Population-Level Effects:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS
## Intercept         0.02     0.01   -0.01    0.04 1.00     5528
## Log10MeanGSH_centered 0.76     0.07    0.62    0.90 1.00     4392
##               Tail_ESS
## Intercept         2718
## Log10MeanGSH_centered 2985
##
## Family Specific Parameters:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma           0.23     0.01    0.21    0.25 1.01     917     1516
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).

```



STAN model

TBD

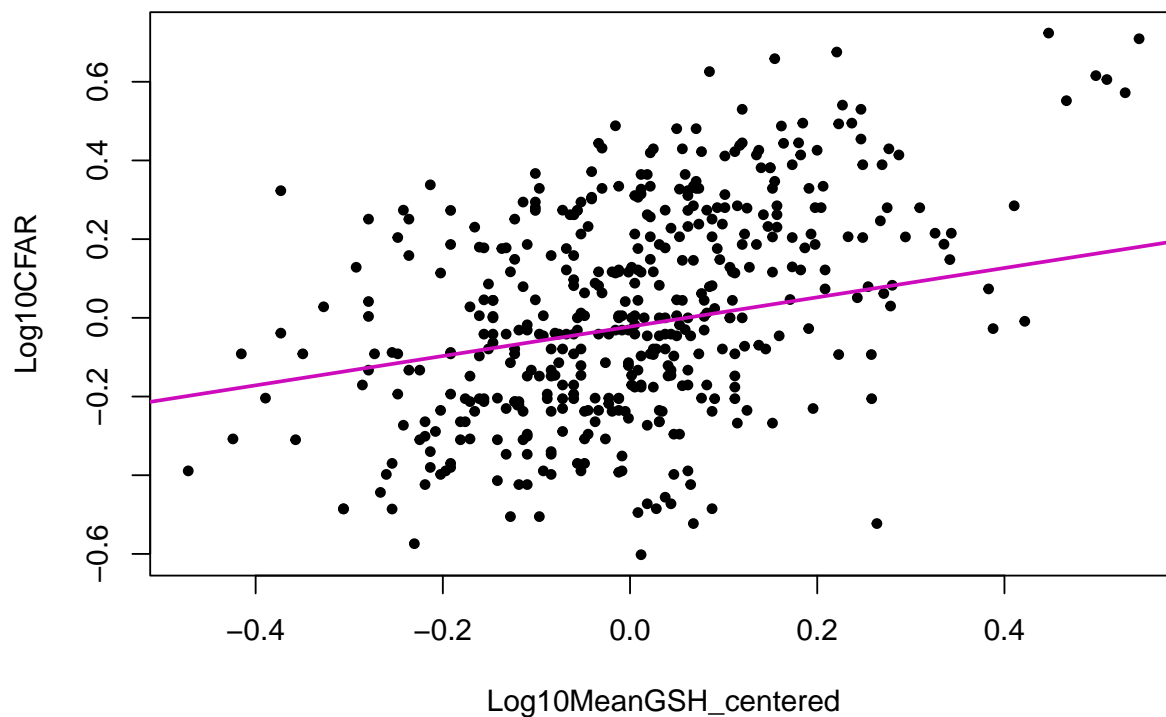
Model 5 - CFAR ~ GSH + Phylogeny

PGLS model

```
pglsMod1 <- gls(Log10CFAR ~ Log10MeanGSH_centered, correlation = corPagel(phy = SampleTree, value = 0.5
```

```
summary(pglsMod1)
```

```
## Generalized least squares fit by maximum likelihood
## Model: Log10CFAR ~ Log10MeanGSH_centered
## Data: GSH_df_pgls
##      AIC      BIC    logLik
## -427.4206 -410.9306 217.7103
##
## Correlation Structure: corPagel
## Formula: ~Binomial
## Parameter estimate(s):
##      lambda
## 0.7030228
##
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept)   -0.0224754 0.06282858 -0.357726  0.7207
## Log10MeanGSH_centered  0.3728837 0.05950418  6.266513  0.0000
##
## Correlation:
##              (Intr)
## Log10MeanGSH_centered -0.065
##
## Standardized residuals:
##      Min      Q1      Med      Q3      Max
## -2.7884571 -0.6933649  0.1151372  1.0885641  2.9034051
##
## Residual standard error: 0.2147249
## Degrees of freedom: 456 total; 454 residual
```



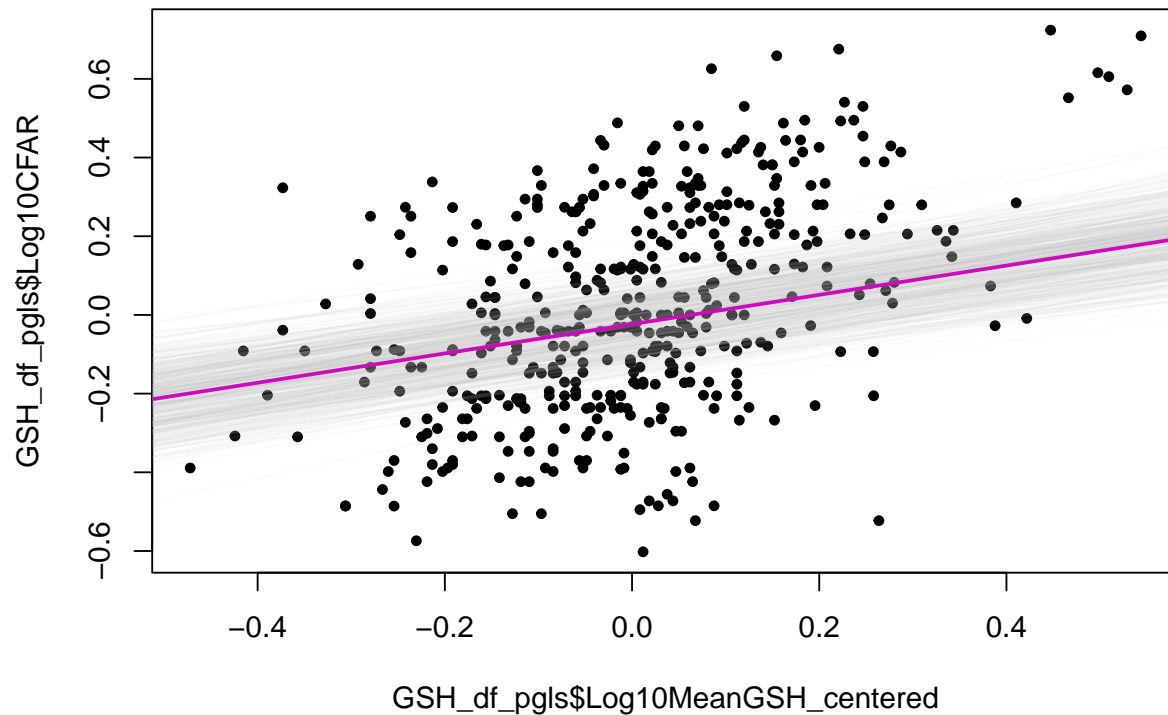
brms model

```
Model_simple <- brm(Log10CFAR ~ Log10MeanGSH_centered + (1|gr(Binomial, cov = A)),
  data = GSH_df_pgls,
  family = gaussian(),
  data2 = list(A = A),
  prior = Simple_prior,
  cores = 2)
```

```
summary(Model_simple)
```

```
## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: Log10CFAR ~ Log10MeanGSH_centered + (1 | gr(Binomial, cov = A))
## Data: GSH_df_pgls (Number of observations: 456)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
## total post-warmup draws = 4000
##
## Group-Level Effects:
## ~Binomial (Number of levels: 456)
## Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept) 0.01 0.00 0.01 0.01 1.00 1200 1938
##
## Population-Level Effects:
```

```
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS
## Intercept          -0.02    0.06   -0.15    0.10 1.00    1257
## Log10MeanGSH_centered  0.37    0.06    0.25    0.49 1.00    4238
##               Tail_ESS
## Intercept           2076
## Log10MeanGSH_centered 3483
##
## Family Specific Parameters:
##               Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma          0.12    0.01    0.11    0.13 1.00    1717    2848
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```



STAN model

Data considered

```
data {
  int <lower=0> N; // number of data points
  int <lower=0> K; // number of predictors
  matrix[N,K] x; // predictor matrix
  vector[N] y; // outcome (CFAR)
```

```

matrix[N, N] d_mat; // sigma matrix
matrix[N, N] A; // vcov matrix
}

```

The parameters accepted by the model.

```

parameters {
  real alpha;
  vector[K] beta; // coefficients
  real<lower=0> sigma; // error
  real<lower=0,upper=1> lambda; // phylogenetic signal
}

transformed parameters {

  matrix[N, N] sigma_mat;
  matrix[N, N] sigma_total;

  vector[N] mu_y;

  sigma_mat = (1-lambda)*d_mat + lambda*A;
  sigma_total = sigma*sigma_mat;

}

```

The model to be estimated. We model the output ‘y’ to be normally distributed with mean ‘mu’ and standard deviation ‘sigma’.

```

model {

  beta ~ student_t(3, 0, 10);
  lambda ~ uniform(0,1);
  sigma ~ student_t(3, 0, 2.5);

  y ~ multi_normal(alpha + x * beta, sigma_total);
}

```

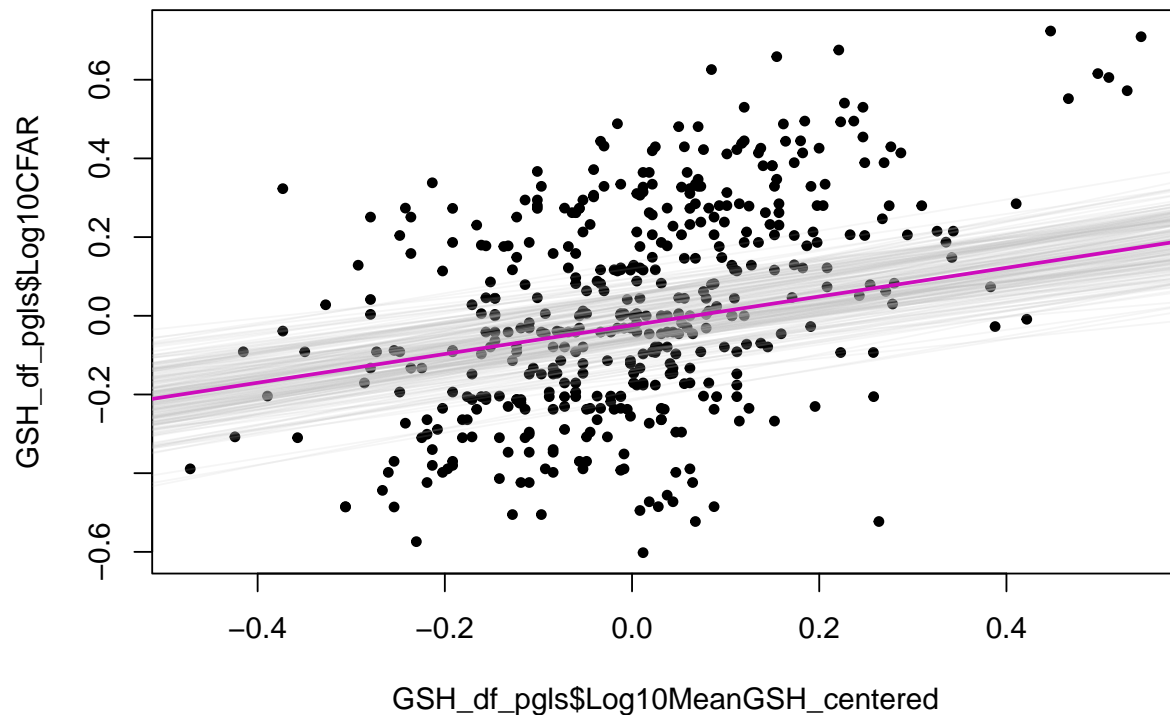
```
fit4
```

```

## Inference for Stan model: StanModel3.
## 4 chains, each with iter=1000; warmup=500; thin=10;
## post-warmup draws per chain=50, total post-warmup draws=200.
##
##           mean se_mean   sd  2.5%   25%   50%   75%  97.5% n_eff Rhat
## alpha    -0.02     0.0 0.06  -0.15  -0.06  -0.03   0.02   0.09   219 0.99
## beta[1]    0.37     0.0 0.06   0.25   0.33   0.36   0.41   0.47   216 1.01
## lp__      625.81     0.1 1.36 622.82 625.03 626.10 626.84 627.47   193 1.02
##

```

```
## Samples were drawn using NUTS(diag_e) at Thu Mar 03 14:30:50 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```



Model 5 - CFAR ~ GSH * Max Size + Phylogeny

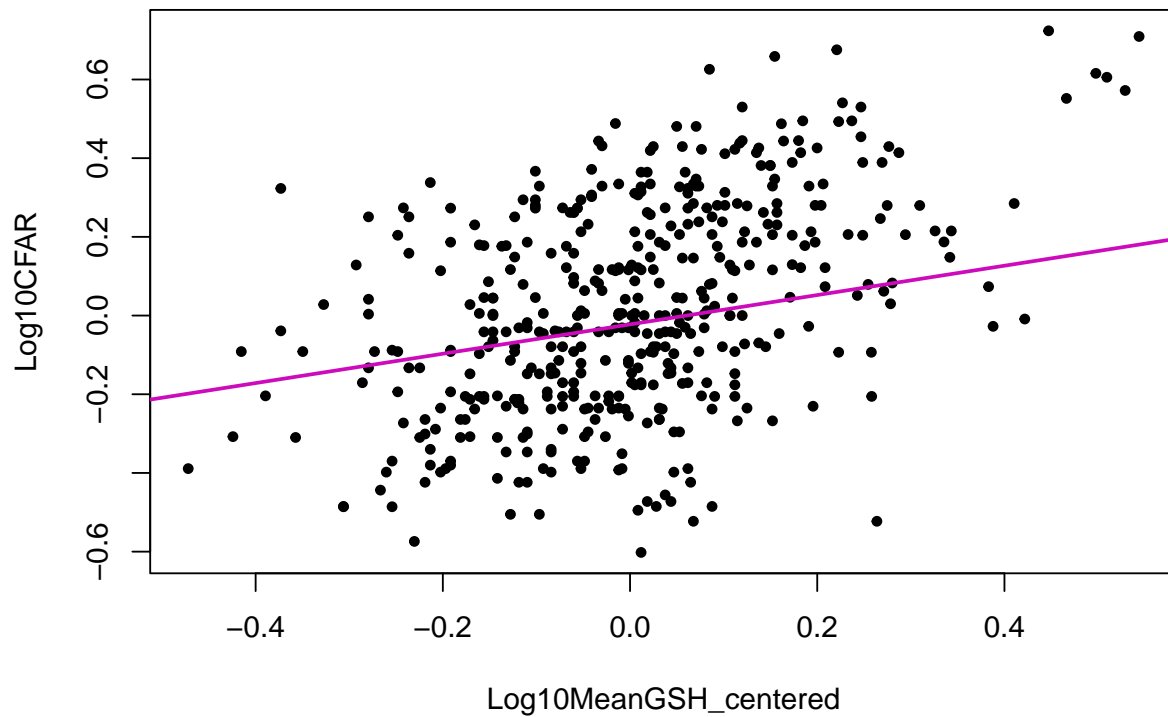
PGLS model

```
pglsMod2 <- gls(Log10CFAR ~ Log10MeanGSH_centered * Log10MaxSize_centered, correlation = corPagel(phy =
```

```
summary(pglsMod2)
```

```
## Generalized least squares fit by maximum likelihood
## Model: y ~ x
## Data: GSH_df_pgl
##      AIC      BIC    logLik
## -427.4206 -410.9306 217.7103
##
## Correlation Structure: corPagel
## Formula: ~Binomial
## Parameter estimate(s):
##      lambda
```

```
## 0.7030228
##
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept) -0.0224754 0.06282858 -0.357726  0.7207
## x            0.3728837 0.05950418  6.266513  0.0000
##
## Correlation:
## (Intr)
## x -0.065
##
## Standardized residuals:
##      Min      Q1      Med      Q3      Max
## -2.7884571 -0.6933649  0.1151372  1.0885641  2.9034051
##
## Residual standard error: 0.2147249
## Degrees of freedom: 456 total; 454 residual
```



brms model

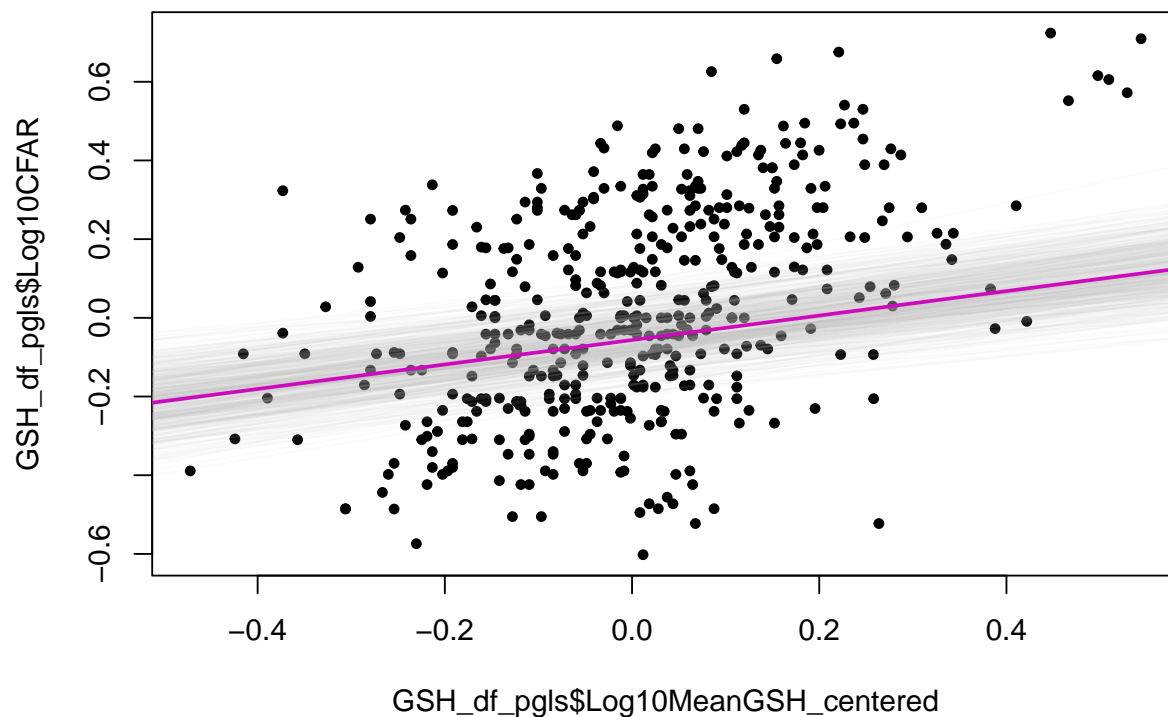
```
Model_BS <- brm(Log10CFAR ~ Log10MeanGSH_centered * Log10MaxSize_centered + (1|gr(Binomial, cov = A)),
  data = GSH_df_pgls,
  family = gaussian(),
  data2 = list(A = A),
```



```
prior = BS_prior,
sample_prior = TRUE, chains = 4, cores = 2)
```

```
summary(Model_BS)
```

```
## Family: gaussian
## Links: mu = identity; sigma = identity
## Formula: Log10CFAR ~ Log10MeanGSH_centered * Log10MaxSize_centered + (1 | gr(Binomial, cov = A))
## Data: GSH_df_pgls (Number of observations: 456)
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
## total post-warmup draws = 4000
##
## Group-Level Effects:
## ~Binomial (Number of levels: 456)
##      Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)      0.01      0.00      0.01      0.01 1.00      1303      2376
##
## Population-Level Effects:
##      Estimate Est.Error 1-95% CI
## Intercept          -0.06      0.06    -0.17
## Log10MeanGSH_centered      0.31      0.06      0.19
## Log10MaxSize_centered      0.13      0.04      0.06
## Log10MeanGSH_centered:Log10MaxSize_centered      0.43      0.12      0.20
##      u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept          0.05 1.00      1824      2402
## Log10MeanGSH_centered      0.43 1.00      6595      3478
## Log10MaxSize_centered      0.20 1.00      4866      3342
## Log10MeanGSH_centered:Log10MaxSize_centered      0.67 1.00      5734      3590
##
## Family Specific Parameters:
##      Estimate Est.Error 1-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma      0.12      0.01      0.11      0.13 1.00      2164      3283
##
## Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
## and Tail_ESS are effective sample size measures, and Rhat is the potential
## scale reduction factor on split chains (at convergence, Rhat = 1).
```



STAN model

Data considered

```
data {
  int <lower=0> N; // number of data points
  int <lower=0> K; // number of predictors
  matrix[N,K] x; // predictor matrix
  vector[N] y; // outcome (CFAR)
  matrix[N, N] d_mat; // sigma matrix
  matrix[N, N] A; // vcov matrix
}
```

The parameters accepted by the model.

```
parameters {
  real alpha;
  vector[K] beta; // coefficients
  real<lower=0> sigma; // error
  real<lower=0,upper=1> lambda; // phylogenetic signal
}

transformed parameters {
```

```

matrix[N, N] sigma_mat;
matrix[N, N] sigma_total;

vector[N] mu_y;

sigma_mat = (1-lambda)*d_mat + lambda*A;
sigma_total = sigma*sigma_mat;

}

```

The model to be estimated. We model the output ‘y’ to be normally distributed with mean ‘mu’ and standard deviation ‘sigma’.

```

model {

  beta ~ student_t(3, 0, 10);
  lambda ~ uniform(0,1);
  sigma ~ student_t(3, 0, 2.5);

  y ~ multi_normal(alpha + x * beta, sigma_total);
}

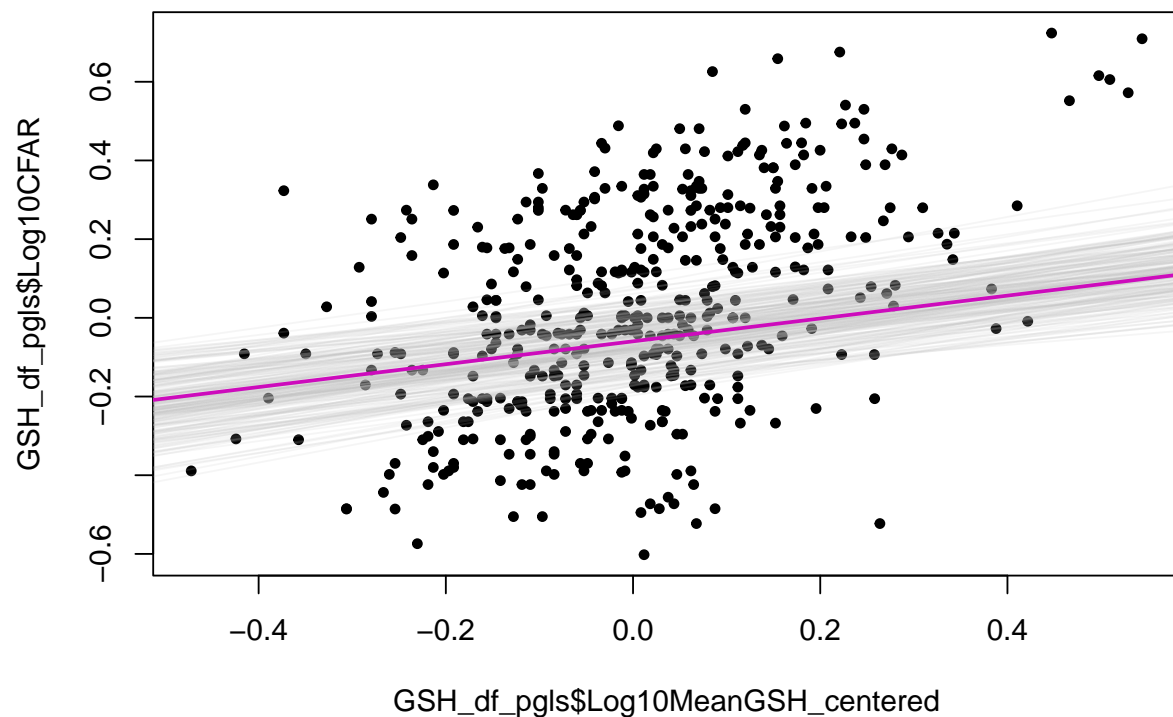
```

```
fit5
```

```

## Inference for Stan model: StanModel3.
## 4 chains, each with iter=1000; warmup=500; thin=10;
## post-warmup draws per chain=50, total post-warmup draws=200.
##
##           mean se_mean   sd  2.5%  25%   50%   75%  97.5% n_eff Rhat
## alpha    -0.06    0.00 0.06  -0.17 -0.10 -0.06 -0.02   0.05  154 1.01
## beta[1]   0.31    0.00 0.06   0.20  0.26  0.32  0.35   0.44  272 0.99
## beta[2]   0.12    0.00 0.04   0.06  0.10  0.12  0.15   0.20  194 0.99
## beta[3]   0.43    0.01 0.12   0.22  0.35  0.44  0.51   0.65  210 0.99
## lp__      641.68    0.12 1.76 637.09 640.81 642.00 643.00 643.99   199 1.00
##
## Samples were drawn using NUTS(diag_e) at Mon Mar 07 12:09:07 2022.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).

```



Model 6 - CFAR ~ GSH + Max Size + Phylogeny

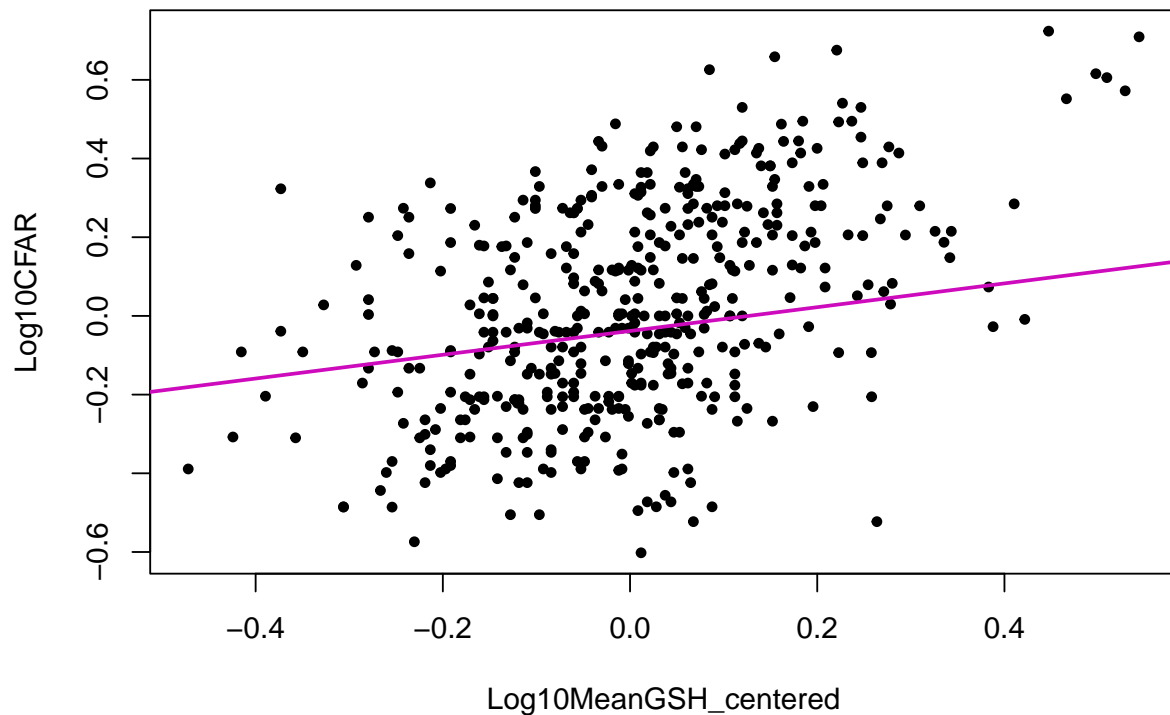
PGLS model

```
pglsMod3 <- gls(Log10CFAR ~ Log10MeanGSH_centered + Log10MaxSize_centered, correlation = corPagel(phy =
```

```
summary(pglsMod3)
```

```
## Generalized least squares fit by maximum likelihood
## Model: Log10CFAR ~ Log10MeanGSH_centered + Log10MaxSize_centered
## Data: GSH_df_pgl
##      AIC      BIC    logLik
## -447.4392 -426.8268 228.7196
##
## Correlation Structure: corPagel
## Formula: ~Binomial
## Parameter estimate(s):
##      lambda
## 0.6786106
##
## Coefficients:
##              Value Std.Error   t-value p-value
## (Intercept) -0.03827951 0.05915461 -0.647110  0.5179
```

```
## Log10MeanGSH_centered  0.30222706 0.06034147  5.008613  0.0000
## Log10MaxSize_centered  0.16260955 0.03412415  4.765233  0.0000
##
## Correlation:
##                      (Intr) L10MGS
## Log10MeanGSH_centered -0.050
## Log10MaxSize_centered -0.056 -0.268
##
## Standardized residuals:
##      Min      Q1      Med      Q3      Max
## -2.8353741 -0.5472167  0.2440250  1.1919410  2.8367324
##
## Residual standard error: 0.2047976
## Degrees of freedom: 456 total; 453 residual
```



brms model

TBD

STAN model

TBD