



BAYESIAN REGRESSION MODELING WITH RSTANARM

Using the R^2 Statistic

Jake Thompson

Psychometrician, ATLAS, University of Kansas



What is the R^2 ?

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$



What is the R^2 ?

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_i \overset{\text{Observed value}}{\boxed{y_i}} - \overset{\text{Predicted value}}{\boxed{\hat{y}_i}}^2}{\sum_i (y_i - \bar{y})^2}$$



What is the R^2 ?

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_i \overset{\text{Observed value}}{\boxed{y_i}} - \overset{\text{Predicted value}}{\boxed{\hat{y}_i}}^2}{\sum_i \overset{\text{Observed value}}{\boxed{y_i}} - \overset{\text{Mean value}}{\boxed{\bar{y}}}^2}$$



Calculating the R^2

```
lm_model <- lm(kid_score ~ mom_iq, data = kidiq)
```

```
lm_summary <- summary(lm_model)
```

```
lm_summary$r.squared
```

```
#> [1] 0.2009512
```

```
ss_res <- var(residuals(lm_model))
```

```
ss_total <- var(residuals(lm_model)) + var(fitted(lm_model))
```

```
1 - (ss_res / ss_total)
```

```
#> [1] 0.2009512
```



The R^2 of a Bayesian Model

```
stan_model <- stan_glm(kid_score ~ mom_iq, data = kidiq)

ss_res <- var(residuals(stan_model))
ss_total <- var(fitted(stan_model)) + var(residuals(stan_model))
1 - (ss_res / ss_total)
#> [1] 0.2004996
```

```
lm_summary$r.squared
#> [1] 0.2009512
```



BAYESIAN REGRESSION MODELING WITH RSTANARM

Let's practice!



BAYESIAN REGRESSION MODELING WITH RSTANARM

Posterior Predictive Model Checks

Jake Thompson

Psychometrician, ATLAS, University of Kansas

Using posterior distributions

```
stan_model <- stan_glm(kid_score ~ mom_iq, data = kidiq)
```

```
spread_draws(stan_model, `(Intercept)`, mom_iq) %>%  
  select(-.draw)  
#> # A tibble: 4,000 x 4  
#>   .chain .iteration `(Intercept)` mom_iq  
#>   <int>   <int>      <dbl>   <dbl>  
#> 1     1     1      19.9    0.654  
#> 2     1     2      20.7    0.643  
#> 3     1     3      27.2    0.604  
#> 4     1     4      24.9    0.613  
#> 5     1     5      26.4    0.610  
#> 6     1     6      25.2    0.619  
#> 7     1     7      17.8    0.702  
#> 8     1     8      35.5    0.502  
#> 9     1     9      32.9    0.540  
#> 10    1    10      27.3    0.599  
#> # ... with 3,990 more rows
```



Posterior predictions

```
predictions <- posterior_linpred(stan_model)

predictions[1:10, 1:5]
#>
#> iterations           1           2           3           4           5
#> [1,] 100.18694 79.04791 96.40964 85.76310 81.30045
#> [2,] 100.24843 82.00786 96.98905 87.80231 83.95155
#> [3,] 100.85608 81.13109 97.33146 87.39709 83.23295
#> [4,] 102.31392 80.81881 98.47300 87.64712 83.10930
#> [5,]  97.25617 81.18278 94.38404 86.28879 82.89553
#> [6,] 100.86263 79.89830 97.11655 86.55800 82.13223
#> [7,]  99.36166 81.10329 96.09910 86.90339 83.04887
#> [8,] 101.13487 80.97878 97.53321 87.38173 83.12658
#> [9,]  98.72686 79.97596 95.37629 85.93252 81.97403
#> [10,] 100.22835 81.04603 96.80069 87.13964 83.09007
```



Comparing score distributions

```
predictions <- posterior_linpred(stan_model)

# First replication
iter1 <- predictions[1,]

# Second replication
iter2 <- predictions[2,]

# Data summaries
summary(kidiq$kid_score)
#>      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    20.0   74.0   90.0   86.8   102.0   144.0

summary(iter1)
#>      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   68.54   79.86   85.80   87.14   93.74   112.12

summary(iter2)
#>      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>   70.05   80.19   85.51   86.71   92.62   109.08
```



Comparing single scores

```
predictions <- posterior_linpred(stan_model)

kidiq$kid_score[24]
#> [1] 87
summary(predictions[, 24])
#>      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
#>  83.34   86.17   86.77   86.75   87.34   90.23
```

```
kidiq$kid_score[185]
#> [1] 111
summary(predictions[, 185])
#>      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
#>  82.81   85.65   86.25   86.24   86.83   89.69
```



BAYESIAN REGRESSION MODELING WITH RSTANARM

Let's practice



BAYESIAN REGRESSION MODELING WITH RSTANARM

Model Fit With Posterior Predictive Model Checks

[Jake Thompson](#)

Psychometrician, ATLAS, University of Kansas



R² Posterior Distribution

```
stan_model <- stan_glm(kid_score ~ mom_iq, data = kidiq)

r2_posterior <- bayes_R2(stan_model)
```

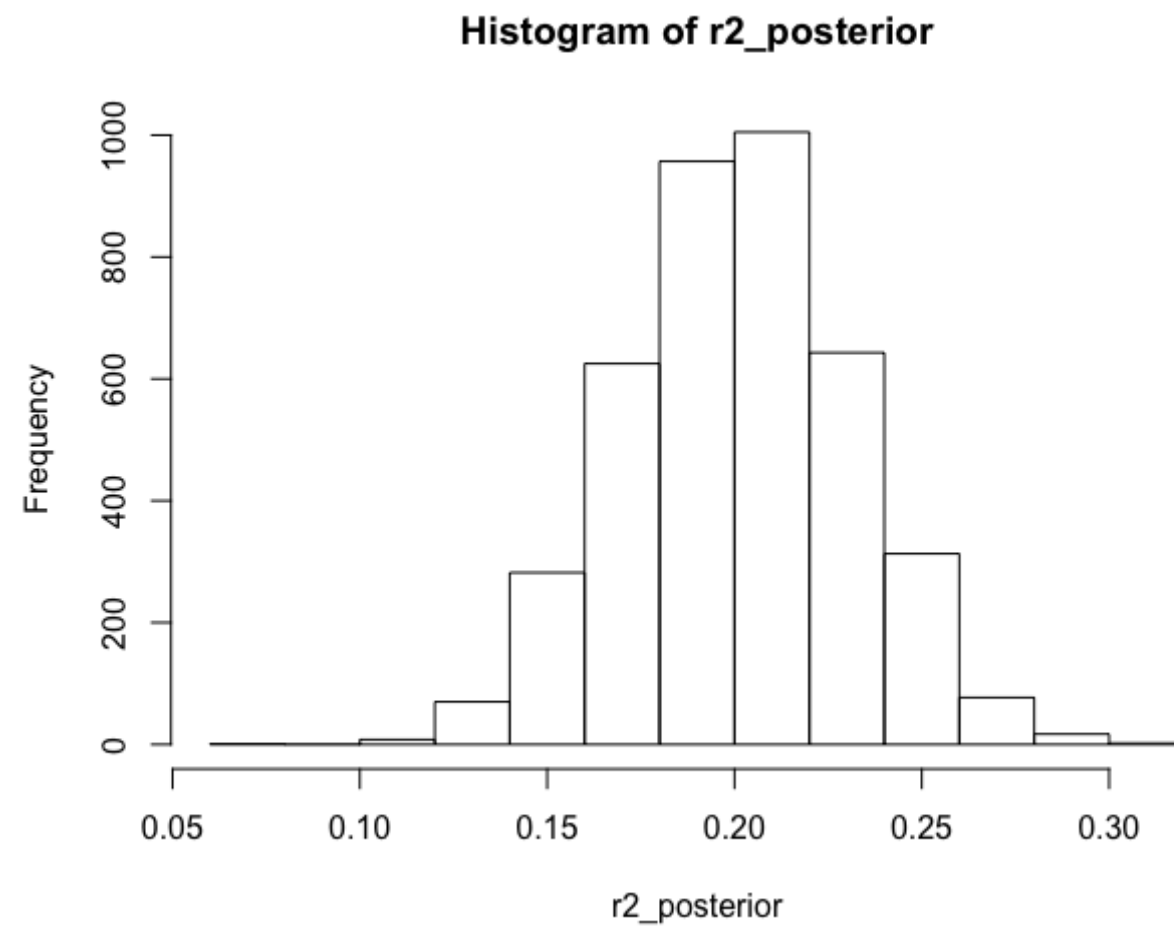
```
summary(r2_posterior)
#>      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#> 0.09677 0.18034 0.20006 0.20042 0.22048 0.33414
```

```
quantile(r2_posterior, probs = c(0.025, 0.975))
#>      2.5%      97.5%
#> 0.1402846 0.2619605
```



R² histogram

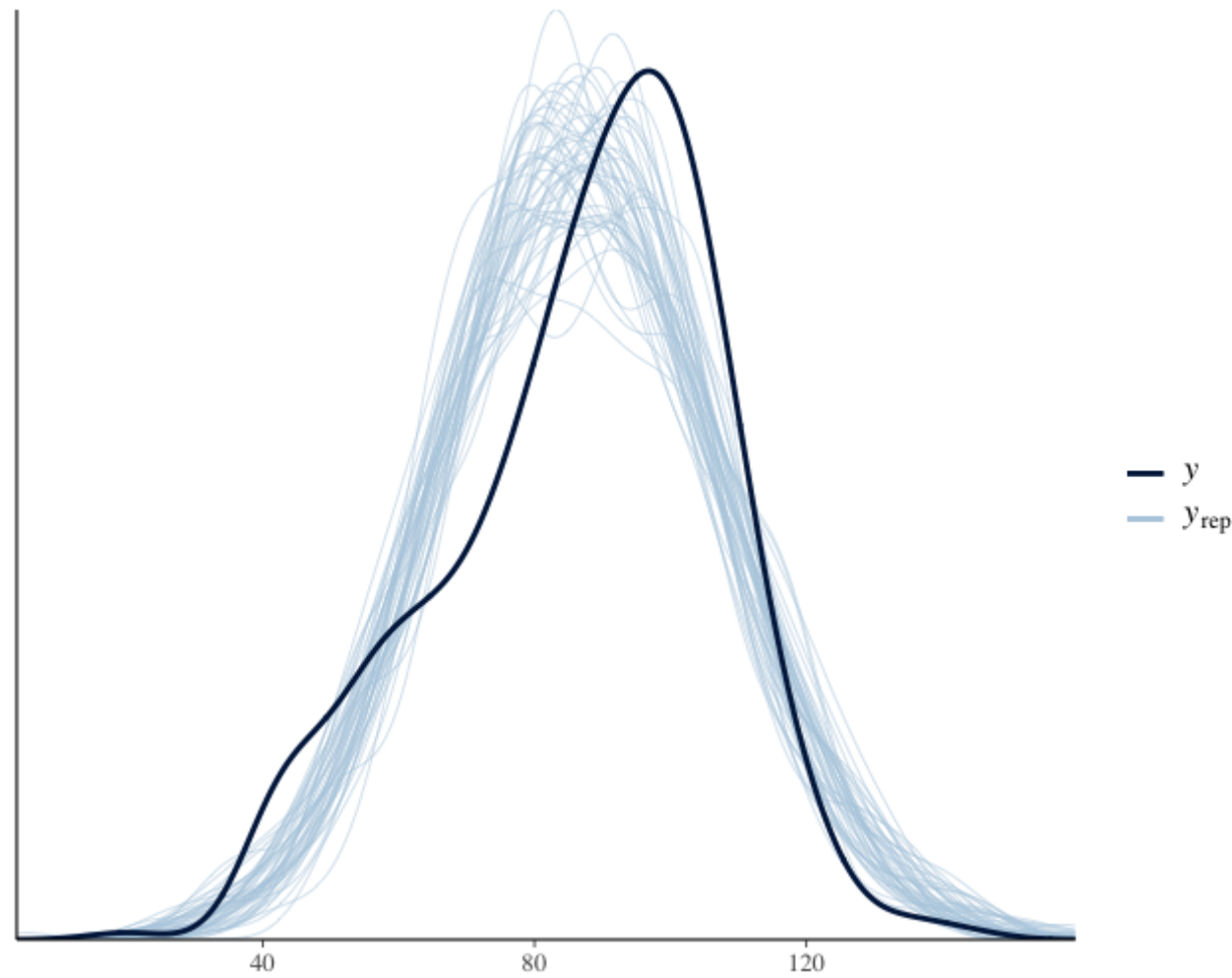
```
hist(r2_posterior)
```





Density Overlay

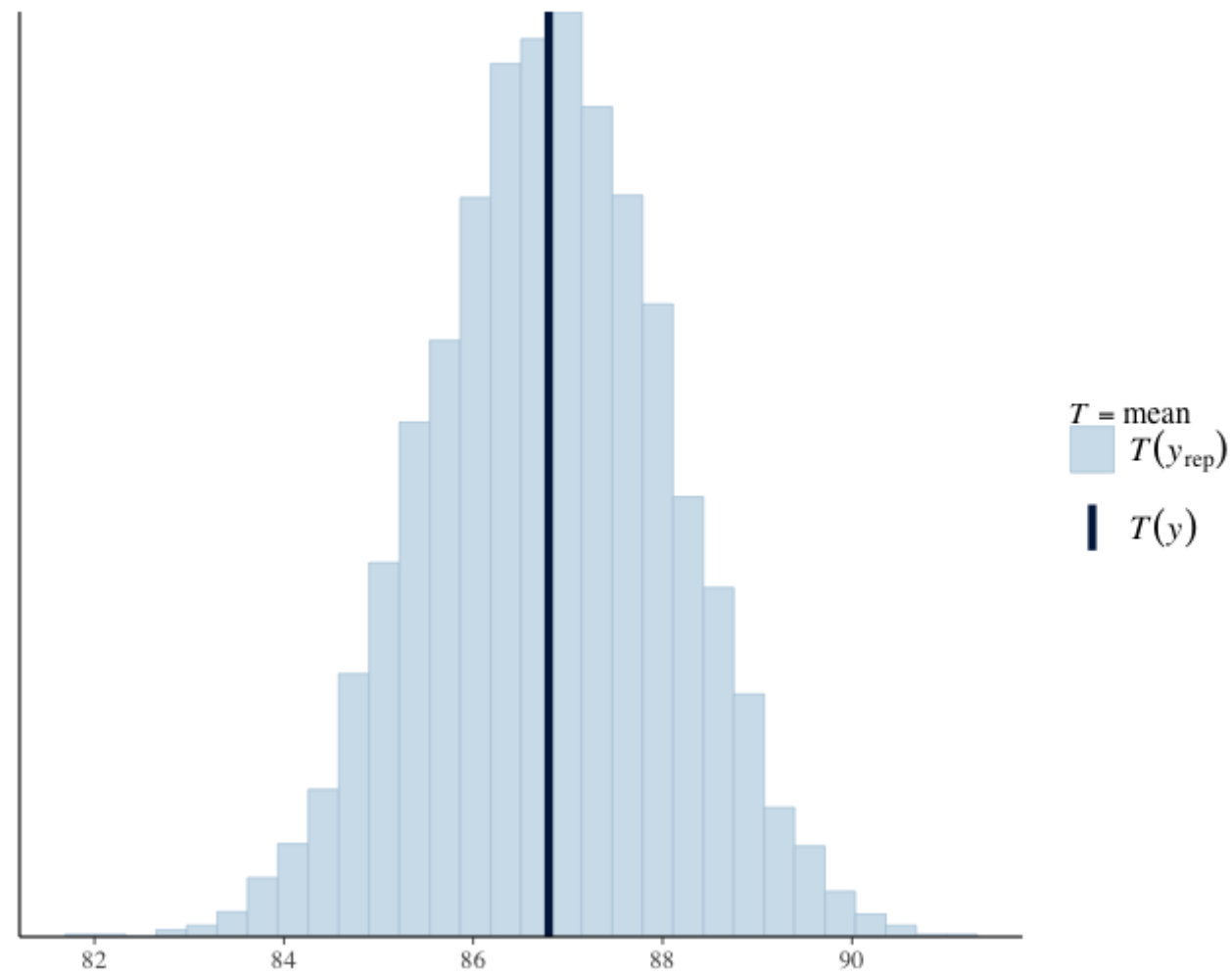
```
pp_check(stan_model, "dens_overlay")
```





Posterior predictive tests

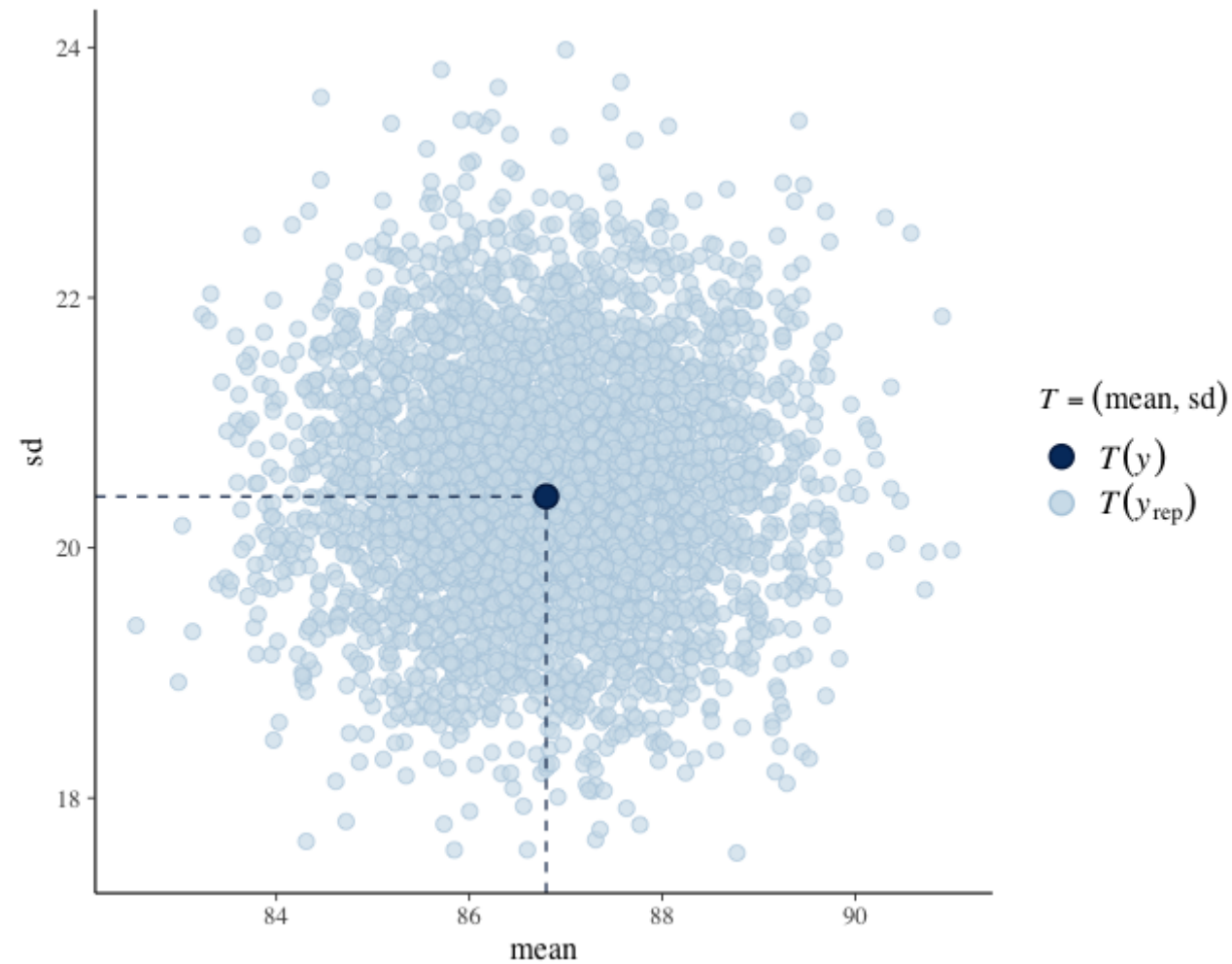
```
pp_check(stan_model, "stat")
```





Posterior predictive tests

```
pp_check(stan_model, "stat_2d")
```





BAYESIAN REGRESSION MODELING WITH RSTANARM

Let's practice!



BAYESIAN REGRESSION MODELING WITH RSTANARM

Bayesian Model Comparisons

Jake Thompson

Psychometrician, ATLAS, University of Kansas



The loo package

- LOO = leave-one-out
 - Approximated cross validation
 - `?loo-package`
 - Using `loo` for model comparisons



Using loo on a single model

```
library(rstanarm)
library(loo)

stan_model <- stan_glm(kid_score ~ mom_iq, data = kidiq)

loo(stan_model)

#>
#> Computed from 4000 by 434 log-likelihood matrix
#>
#>           Estimate      SE
#> elpd_loo  -1878.5  14.5
#> p_loo       2.9   0.3
#> looic       3757.1 29.0
#> -----
#> Monte Carlo SE of elpd_loo is 0.0.
#>
#> All Pareto k estimates are good (k < 0.5).
#> See help('pareto-k-diagnostic') for details.
```



Model comparisons with loo

```
model_1pred <- stan_glm(kid_score ~ mom_iq, data = kidiq)
model_2pred <- stan_glm(kid_score ~ mom_iq * mom_hs, data = kidiq)

loo_1pred <- loo(model_1pred)
loo_2pred <- loo(model_2pred)

compare(loo_1pred, loo_2pred)
#>   elpd_diff      se
#>      6.1      3.9
```




Model comparisons with loo

```
compare(loo_1pred, loo_2pred)
#> elpd_diff      se
#>      6.1      3.9
```

- Positive = prefer second model
- Negative = prefer first model
- Significant difference?
 - Absolute value of difference relative to standard error



BAYESIAN REGRESSION MODELING WITH RSTANARM

Let's practice!