# Variable selection with projpred

- In your project it is sufficient to compare 2–3 models

# Variable selection with projpred

- In your project it is sufficient to compare 2–3 models
- ...but if you are interested in variable selection, then the number of potential models is $2^p$, where $p$ is the number of variables

# Variable selection with projpred

- In your project it is sufficient to compare 2–3 models
- ...but if you are interested in variable selection, then the number of potential models is $2^p$, where $p$ is the number of variables
- ...in such case I recommended to use `brms` + `projpred`

# Variable selection with projpred

- In your project it is sufficient to compare 2–3 models
- ...but if you are interested in variable selection, then the number of potential models is $2^p$, where $p$ is the number of variables
- ...in such case I recommended to use `brms` + `projpred`
- `projpred` avoids the overfit in model selection

# Use of reference models in model selection

- Background
- First example
- Bayesian and decision theoretical justification
- More examples

# Not a novel idea

- Lindley (1968): *The choice of variables in multiple regression*
    - Bayesian and decision theoretical justification, but simplified model and computation

# Not a novel idea

- Lindley (1968): *The choice of variables in multiple regression*
  - Bayesian and decision theoretical justification, but simplified model and computation
- Goutis & Robert (1998): *Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections*
  - one key part for practical computation

# Not a novel idea

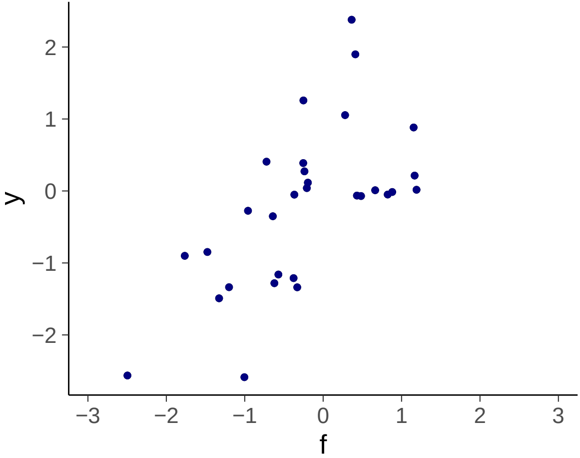
- Lindley (1968): *The choice of variables in multiple regression*
  - Bayesian and decision theoretical justification, but simplified model and computation
- Goutis & Robert (1998): *Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections*
  - one key part for practical computation
- Related approaches
  - gold standard, preconditioning, teacher and student, distilling, . . .

# Not a novel idea

- Lindley (1968): *The choice of variables in multiple regression*
  - Bayesian and decision theoretical justification, but simplified model and computation
- Goutis & Robert (1998): *Model choice in generalised linear models: a Bayesian approach via Kullback-Leibler projections*
  - one key part for practical computation
- Related approaches
  - gold standard, preconditioning, teacher and student, distilling, . . .
- Motivation in these
  - measurement cost in covariates
  - running cost of predictive model
  - easier explanation / learn from the model

# Example: Simulated regression

$$
\begin{aligned}
f &\sim \mathrm{N}(0, 1), \\
y \mid f &\sim \mathrm{N}(f, 1)
\end{aligned}
$$



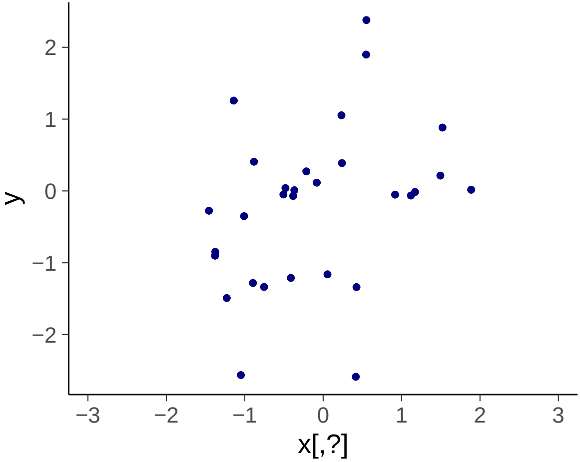
# Example: Simulated regression

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j &= 1,\ldots,150\,, \\
y \mid f &\sim \mathrm{N}(f,1) & x_j \mid f &\sim \mathrm{N}(0,1), & j &= 151,\ldots,500\,.
\end{aligned}
$$

# Example: Simulated regression

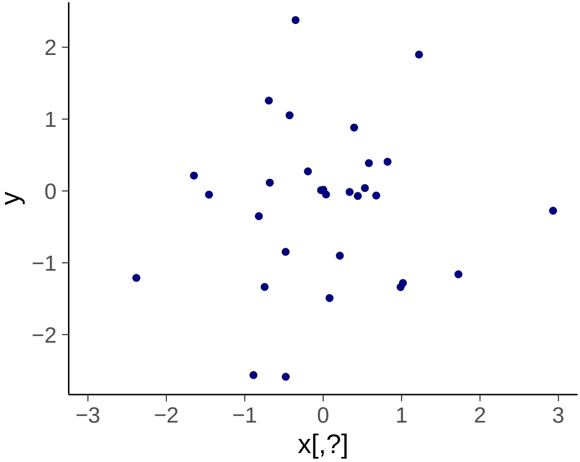
$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$

$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



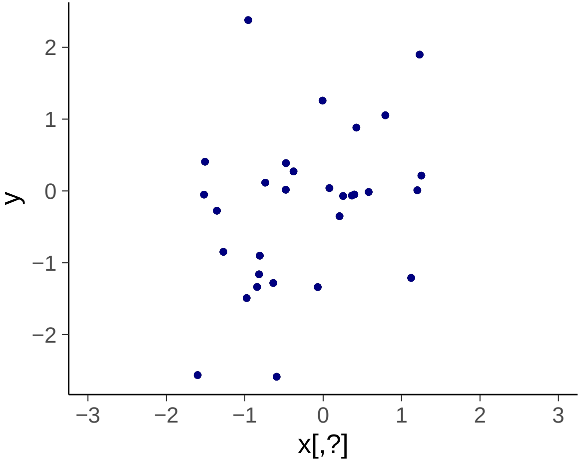
# Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



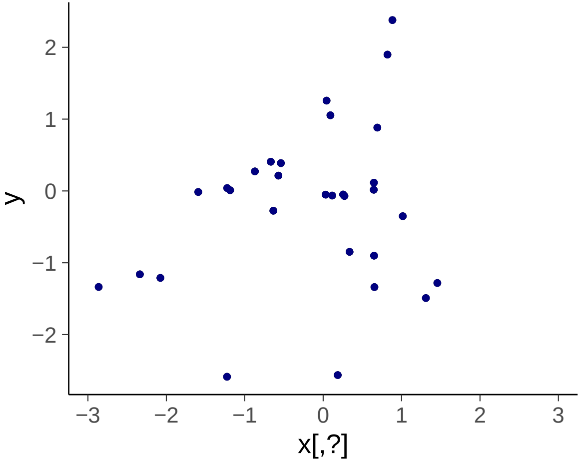
# Example: Simulated regression

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j &= 1,\ldots,150\,, \\
y \mid f &\sim \mathrm{N}(f,1) & x_j \mid f &\sim \mathrm{N}(0,1), & j &= 151,\ldots,500\,.
\end{aligned}
$$



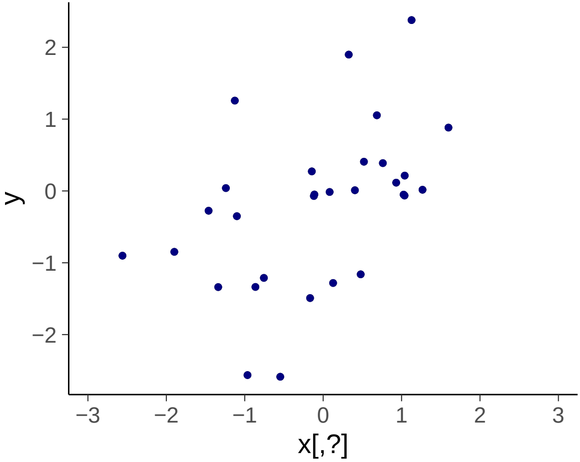
# Example: Simulated regression

$$
\begin{aligned}
&f \sim \mathrm{N}(0,1), && x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), && j = 1, \ldots, 150\,, \\
&y \mid f \sim \mathrm{N}(f, 1) && x_j \mid f \sim \mathrm{N}(0,1), && j = 151, \ldots, 500\,.
\end{aligned}
$$



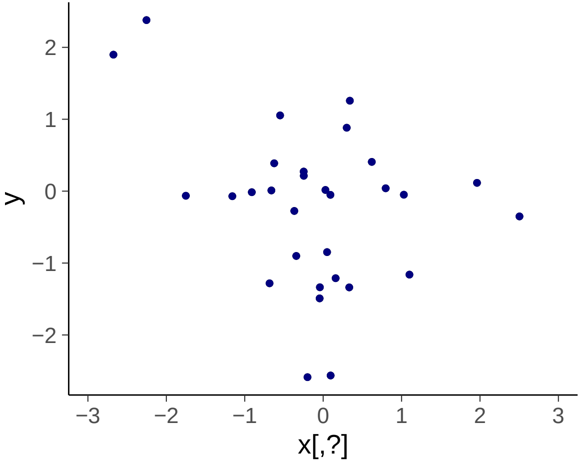
# Example: Simulated regression

$$
\begin{aligned}
&f \sim \mathrm{N}(0,1), && x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), && j = 1, \ldots, 150\,, \\
&y \mid f \sim \mathrm{N}(f,1) && x_j \mid f \sim \mathrm{N}(0,1), && j = 151, \ldots, 500\,.
\end{aligned}
$$



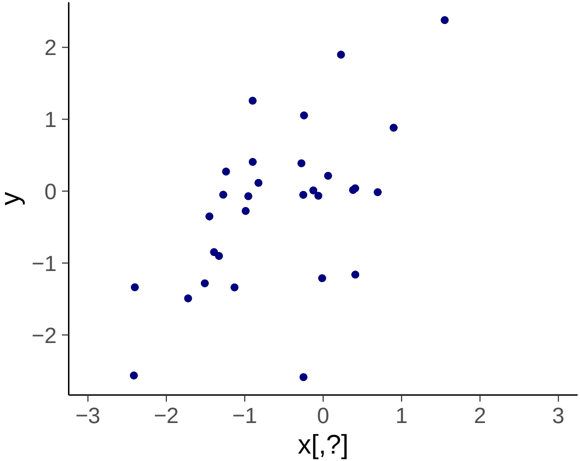
# Example: Simulated regression

$$
\begin{aligned}
&f \sim \mathrm{N}(0,1), \qquad && x_j \mid f \sim \mathrm{N}(\sqrt{\rho}f,\ 1-\rho), \qquad && j = 1, \ldots, 150\,, \\
&y \mid f \sim \mathrm{N}(f,1) \qquad && x_j \mid f \sim \mathrm{N}(0,1), \qquad && j = 151, \ldots, 500\,.
\end{aligned}
$$



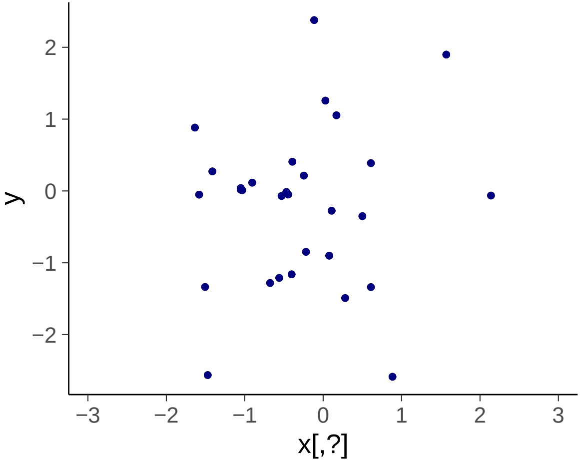
# Example: Simulated regression

$$
\begin{aligned}
&f \sim \mathrm{N}(0,1), && x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), && j = 1,\ldots,150\,, \\
&y \mid f \sim \mathrm{N}(f,1) && x_j \mid f \sim \mathrm{N}(0,1), && j = 151,\ldots,500\,.
\end{aligned}
$$



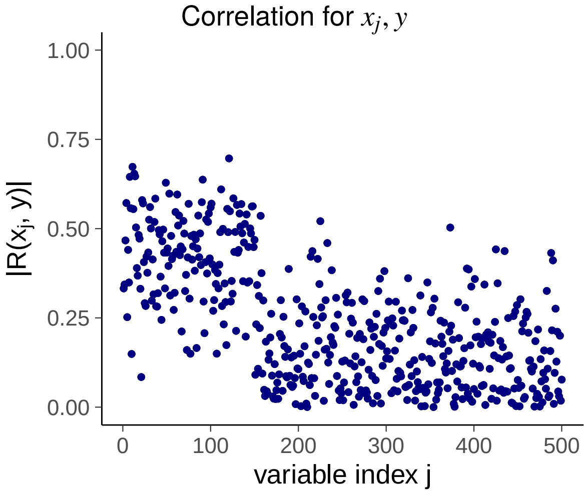
# Example: Simulated regression

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



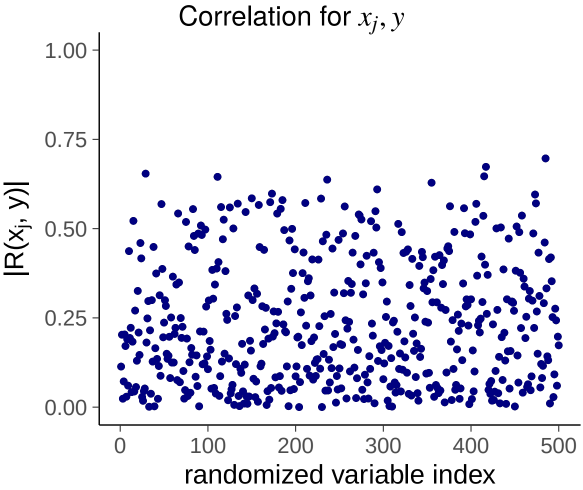
# Example: Individual correlations

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\, 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$



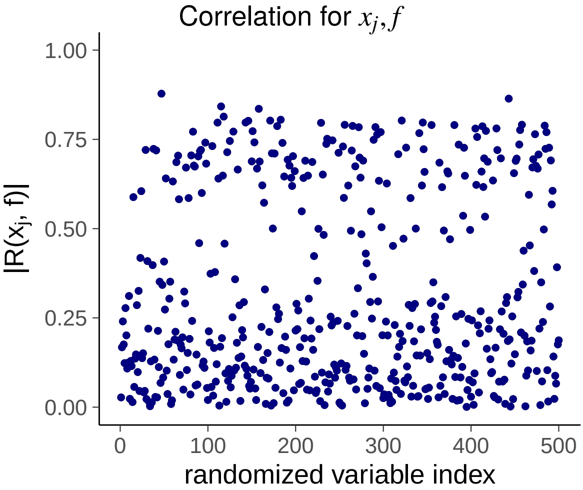
# Example: Individual correlations

$$
\begin{aligned}
f &\sim \mathrm{N}(0,1), & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j &= 1,\ldots,150\,, \\
y \mid f &\sim \mathrm{N}(f,1) & x_j \mid f &\sim \mathrm{N}(0,1), & j &= 151,\ldots,500\,.
\end{aligned}
$$



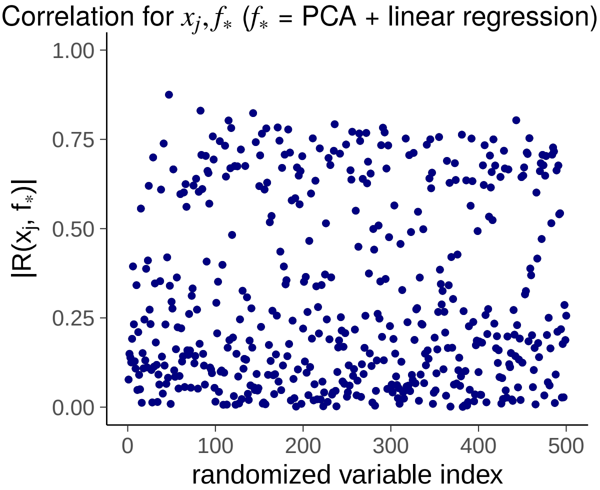
# Example: Individual correlations

$$f \sim \mathrm{N}(0,1), \qquad x_j \mid f \sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), \qquad j = 1, \ldots, 150\,,$$
$$y \mid f \sim \mathrm{N}(f,1) \qquad x_j \mid f \sim \mathrm{N}(0,1), \qquad j = 151, \ldots, 500\,.$$

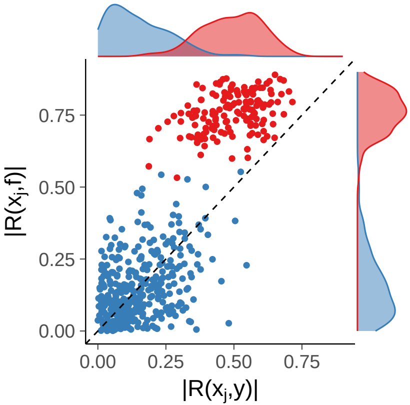


# Example: Individual correlations

$$\begin{aligned} f &\sim \mathrm{N}(0,1), & x_j \mid f &\sim \mathrm{N}(\sqrt{\rho} f,\ 1-\rho), & j &= 1,\ldots,150\,, \\ y \mid f &\sim \mathrm{N}(f,1) & x_j \mid f &\sim \mathrm{N}(0,1), & j &= 151,\ldots,500\,. \end{aligned}$$



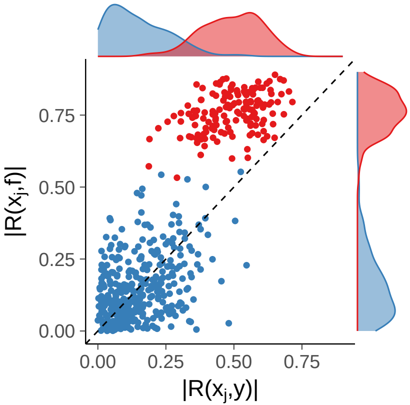
# Knowing the latent values would help

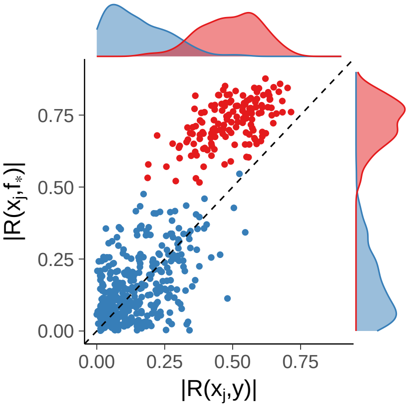



irrelevant $x_j$, relevant $x_j$

A) Sample correlation with $y$ vs. sample correlation with $f$

# Estimating the latent values with a reference model helps




irrelevant $x_j$, relevant $x_j$

A) Sample correlation with $y$ vs. sample correlation with $f$

B) Sample correlation with $y$ vs. sample correlation with $f_*$

$f_* =$ linear regression fit with 3 principal components

# Bayesian justification

- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model

# Bayesian justification

- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model
- Consider model selection as decision problem

# Bayesian justification

- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model
- Consider model selection as decision problem
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible

# Bayesian justification

- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model
- Consider model selection as decision problem
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some $\theta_0$
    $\Rightarrow$ "Optimal point estimates"

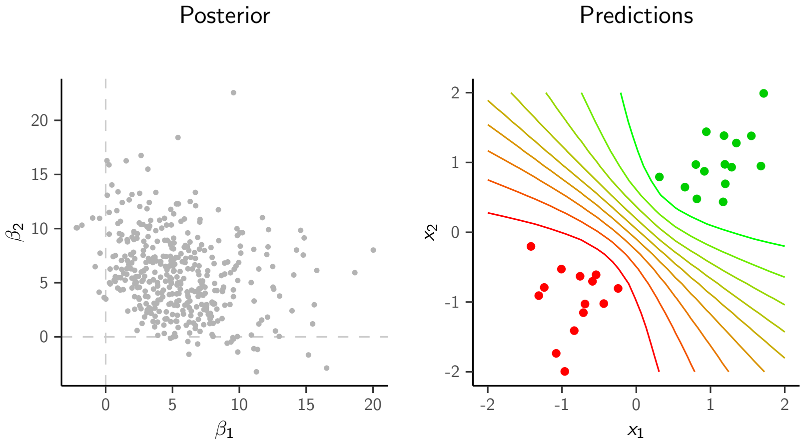
# Bayesian justification

- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model
- Consider model selection as decision problem
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some $\theta_0$
    $\Rightarrow$ "Optimal point estimates"
  - Some covariates must have exactly zero regression coefficient
    $\Rightarrow$ "Which covariates can be discarded"

# Bayesian justification

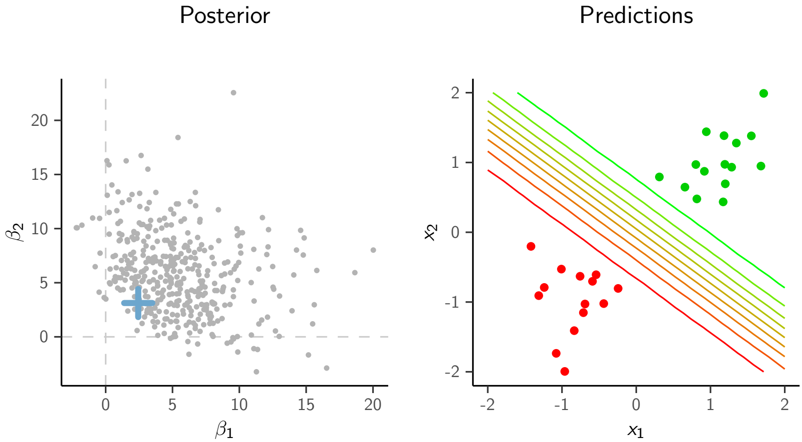
- Theory says to integrate over all the uncertainties
  - build a rich model
  - make model checking etc.
  - this model can be the reference model
- Consider model selection as decision problem
- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- Example constraints
  - $q(\theta)$ can have only point mass at some $\theta_0$
    $\Rightarrow$ "Optimal point estimates"
  - Some covariates must have exactly zero regression coefficient
    $\Rightarrow$ "Which covariates can be discarded"
  - Much simpler model
    $\Rightarrow$ "Easier explanation"

# Logistic regression with two covariates



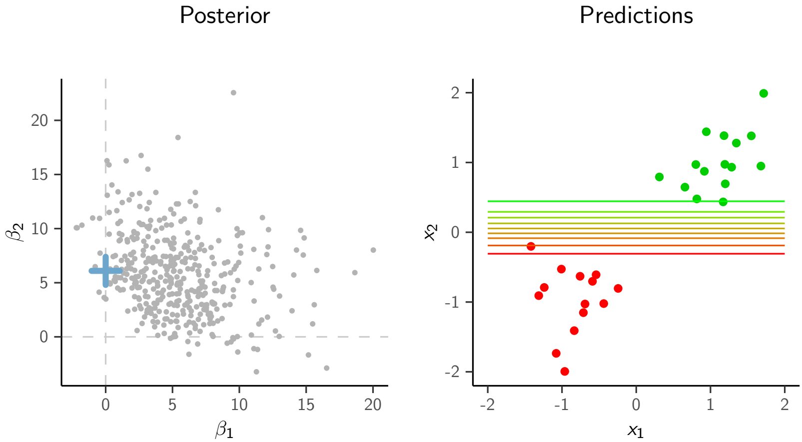

Full posterior for $\beta_1$ and $\beta_2$ and contours of predicted class probability

# Logistic regression with two covariates



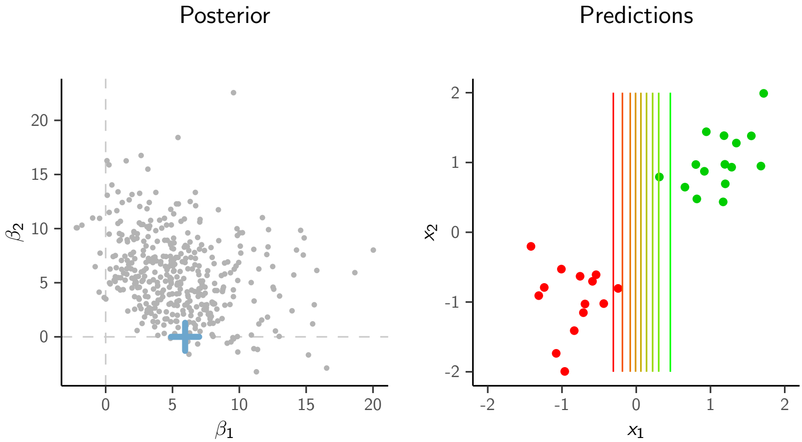

Projected point estimates for $\beta_1$ and $\beta_2$

# Logistic regression with two covariates



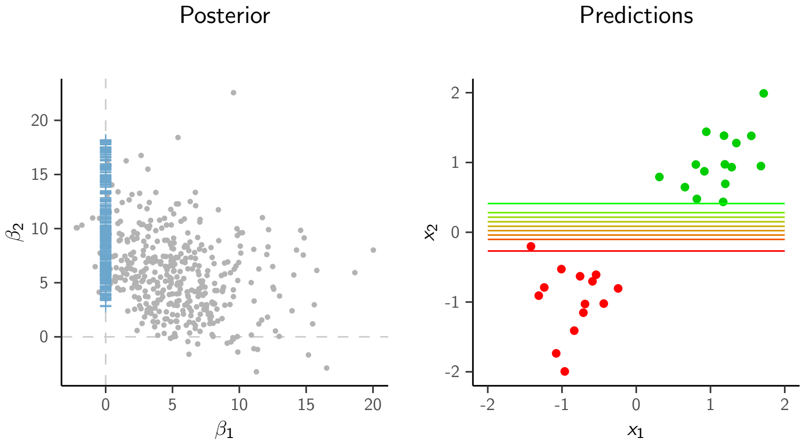

Projected point estimates, constraint $\beta_1 = 0$

# Logistic regression with two covariates



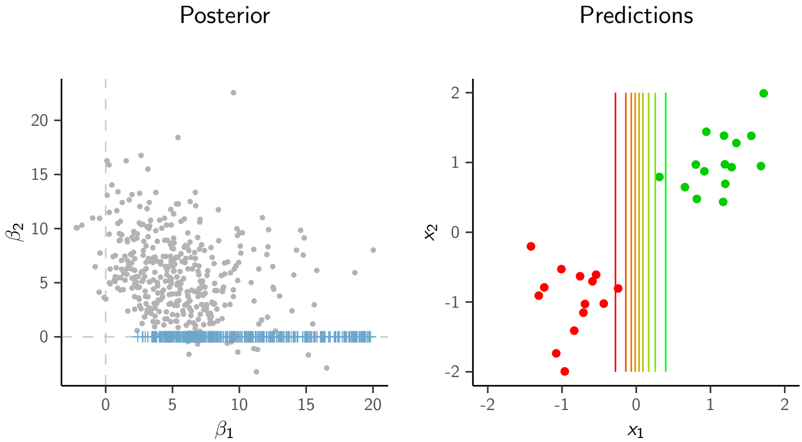

Projected point estimates, constraint $\beta_2 = 0$

# Logistic regression with two covariates




Draw-by-draw projection, constraint $\beta_1 = 0$

# Logistic regression with two covariates




Draw-by-draw projection, constraint $\beta_2 = 0$

# Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible

# Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- As the full posterior $p(\theta \mid D)$ is projected to $q(\theta)$
  - the prior is also projected and there is no need to define priors for submodels separately

# Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- As the full posterior $p(\theta \mid D)$ is projected to $q(\theta)$
  - the prior is also projected and there is no need to define priors for submodels separately
  - even if we constrain some coefficients to be $0$, the predictive inference is conditoned on the information related features contributed to the reference model

# Predictive projection

- Replace full posterior $p(\theta \mid D)$ with some constrained $q(\theta)$ so that the *predictive distribution* changes as little as possible
- As the full posterior $p(\theta \mid D)$ is projected to $q(\theta)$
  - the prior is also projected and there is no need to define priors for submodels separately
  - even if we constrain some coefficients to be $0$, the predictive inference is conditoned on the information related features contributed to the reference model
  - solves the problem of how to do the inference after the model selection

# Projective selection

- How to select a feature combination?

# Projective selection

- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss

# Projective selection

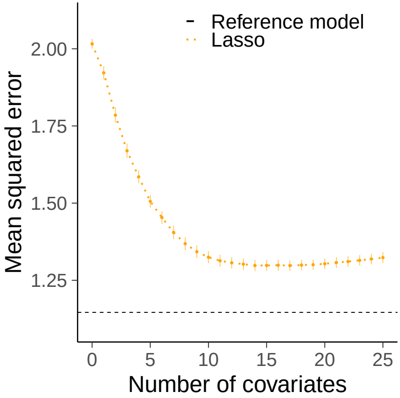
- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss
- Search heuristics, e.g.
  - Monte Carlo search
  - Forward search
  - $L_1$-penalization (as in Lasso)

# Projective selection

- How to select a feature combination?
- For a given model size, choose feature combination with minimal projective loss
- Search heuristics, e.g.
  - Monte Carlo search
  - Forward search
  - $L_1$-penalization (as in Lasso)
- Use cross-validation to select the appropriate model size
  - need to cross-validate over the search paths

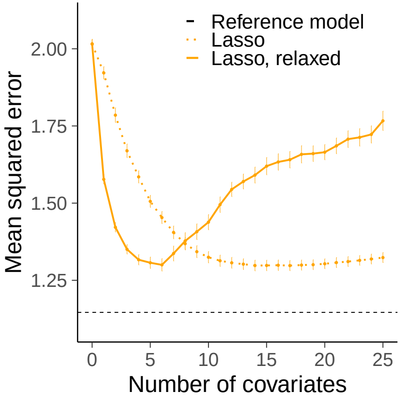
# Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



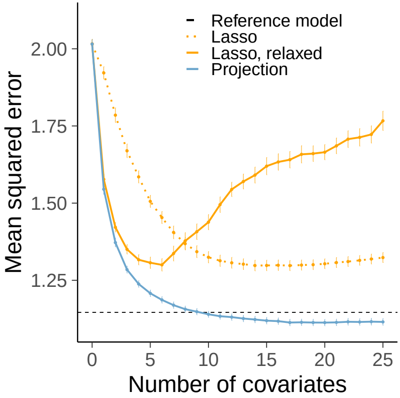
# Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



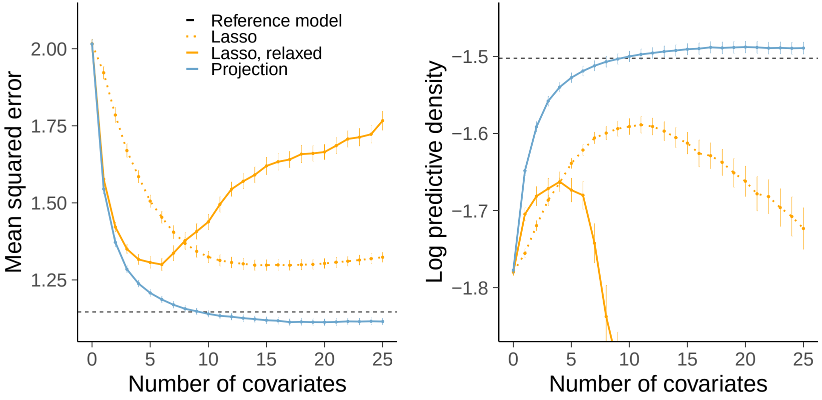
# Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$



# Projective selection vs. Lasso

Same simulated regression data as before,
$n = 50$, $p = 500$, $p_{\text{rel}} = 150$, $\rho = 0.5$

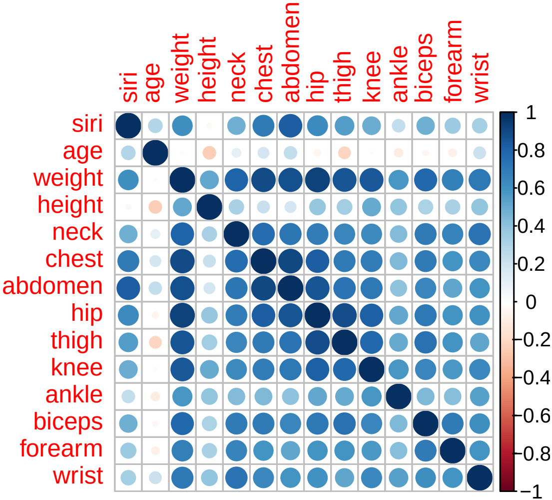


# Bodyfat: small $p$ example of projection predictive

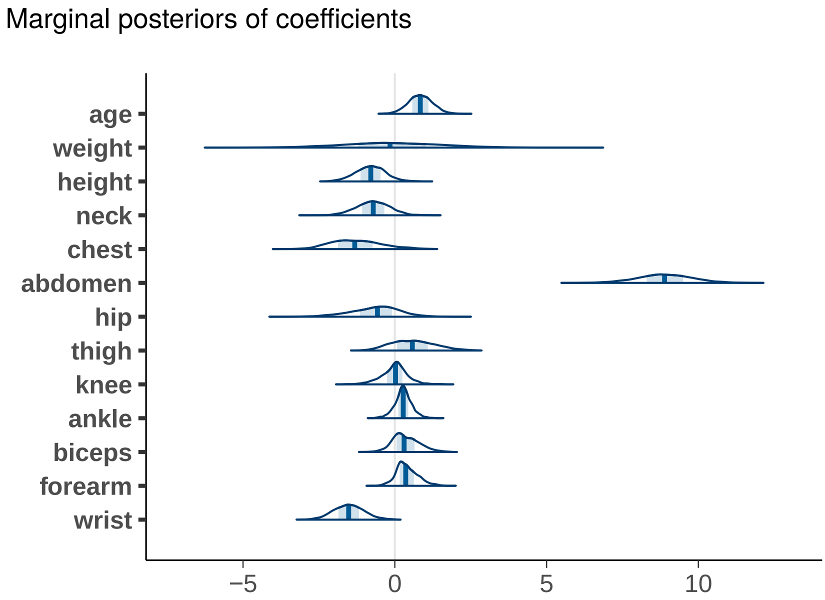
Predict bodyfat percentage. The reference value is obtained by immersing person in water. $n = 251$.

# Bodyfat: small $p$ example of projection predictive

Predict bodyfat percentage. The reference value is obtained by immersing person in water. $n = 251$.

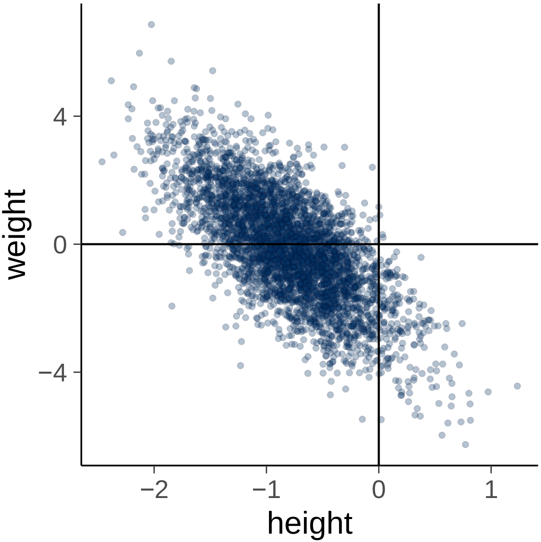


# Bodyfat

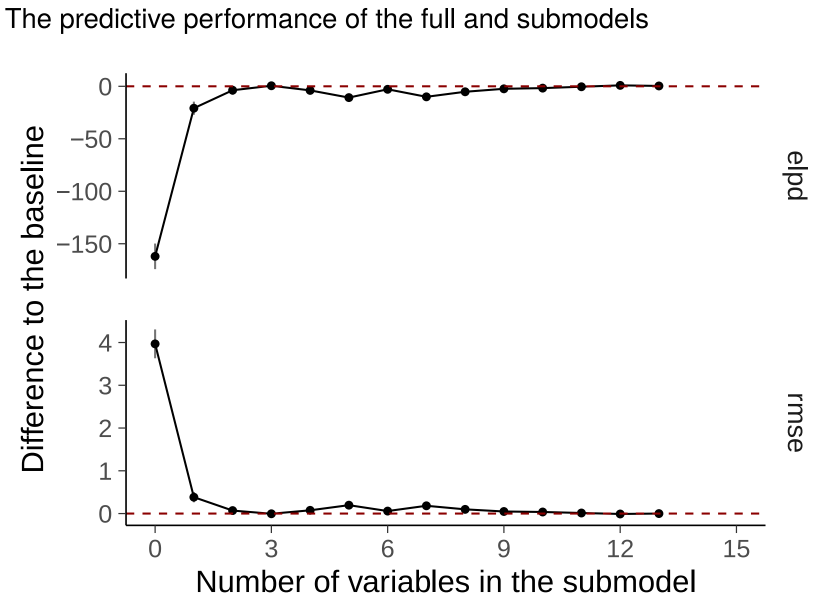

Marginal posteriors of coefficients
age
weight
height
neck
chest
abdomen
hip
thigh
knee
ankle
biceps
forearm
wrist
−5
0
5
10

# Bodyfat

Bivariate marginal of weight and height

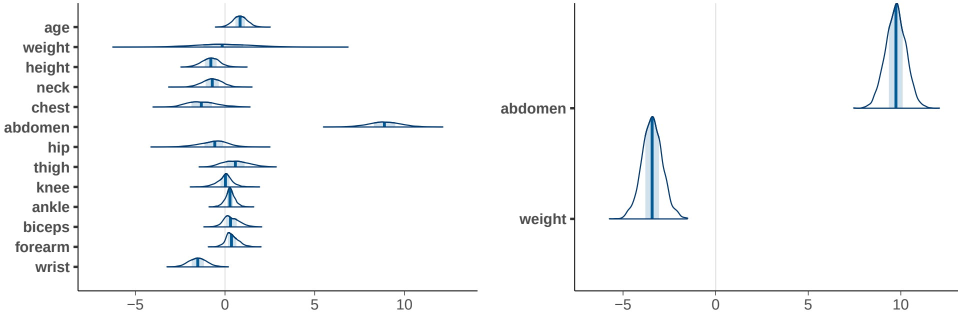


# Bodyfat


The predictive performance of the full and submodels
Difference to the baseline
0
−50
−100
−150
elpd
4
3
2
1
0
rmse
0
3
6
9
12
15
Number of variables in the submodel

# Bodyfat

## Marginals of the reference and projected posterior



# Predictive performance vs. selected variables

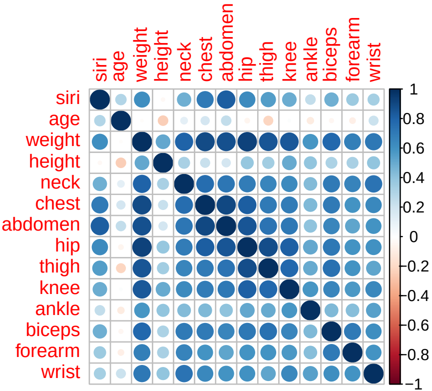
- The initial aim: find the minimal set of variables providing similar predictive performance as the reference model

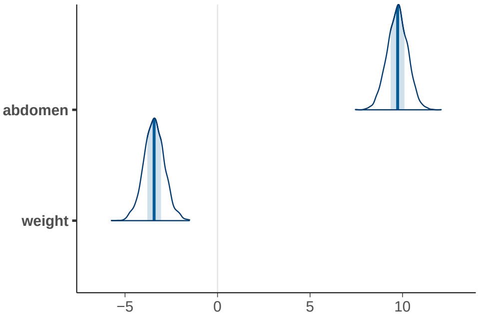
# Predictive performance vs. selected variables

- The initial aim: find the minimal set of variables providing similar predictive performance as the reference model
- Some keep asking can it find the true variables

# Predictive performance vs. selected variables

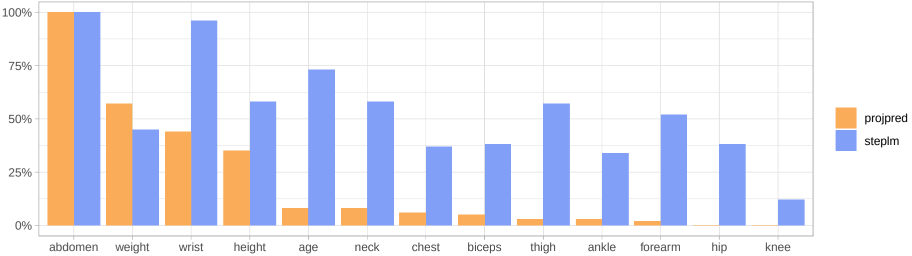
- The initial aim: find the minimal set of variables providing similar predictive performance as the reference model
- Some keep asking can it find the true variables
    - What do you mean by true variables?



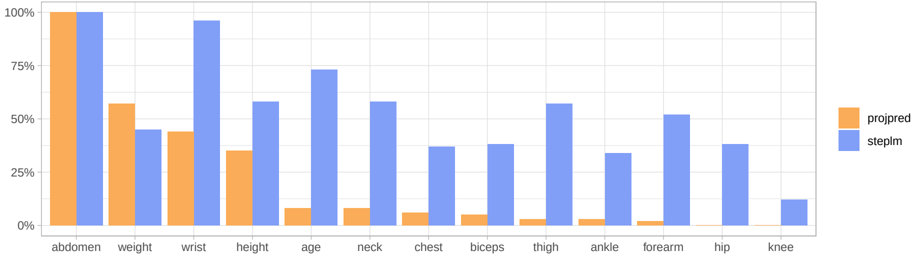
# Variability under data perturbation

Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets



# Variability under data perturbation

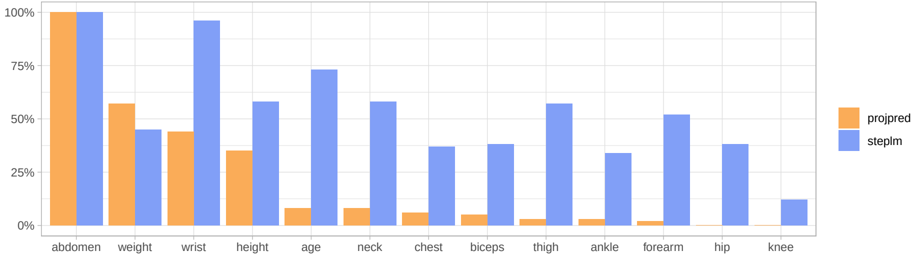
Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets




| M | projpred | Freq % | steplm | Freq % |
|---|---|---|---|---|
| 1 | abdom., weight | 39 | abdom., age, forearm, height, hip, neck, thigh, wrist | 4 |
| 2 | abdom., wrist | 10 | abdom., age, chest, forearm, height, neck, thigh, wrist | 4 |
| 3 | abdom., height | 10 | abdom., forearm, height, neck, wrist | 2 |
| 4 | abdom., height, wrist | 9 | abdom., forearm, neck, weight, wrist | 2 |
| 5 | abdom., weight, wrist | 8 | abdom., age, height, hip, thigh, wrist | 2 |
| 6 | abdom., chest, height, wrist | 2 | abdom., age, height, hip, neck, thigh, wrist | 2 |
| 7 | abdom., biceps, weight, wrist | 2 | abdom., age, ankle, forearm, height, hip, neck, thigh, wrist | 2 |
| 8 | abdom., height, weight, wrist | 2 | abdom., age, biceps, chest, height, neck, wrist | 2 |
| 9 | abdom., age, wrist | 2 | abdom., age, biceps, chest, forearm, height, neck, thigh, wrist | 2 |
| 10 | abdom., age, height, neck, thigh, wrist | 2 | abdom., age, ankle, biceps, weight, wrist | 2 |

# Variability under data perturbation

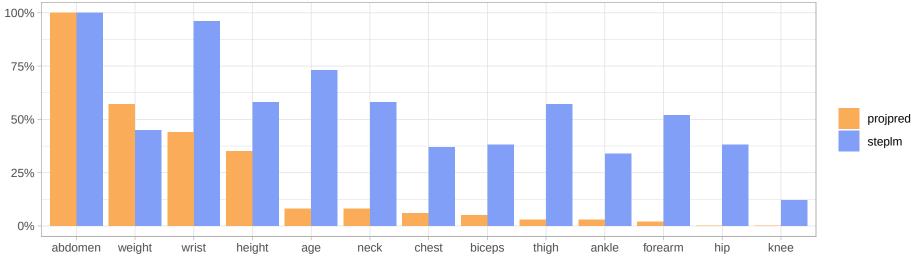
Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets




- Reduced variability, but in case of noisy finite data, there will be some variability under data perturbation

# Variability under data perturbation

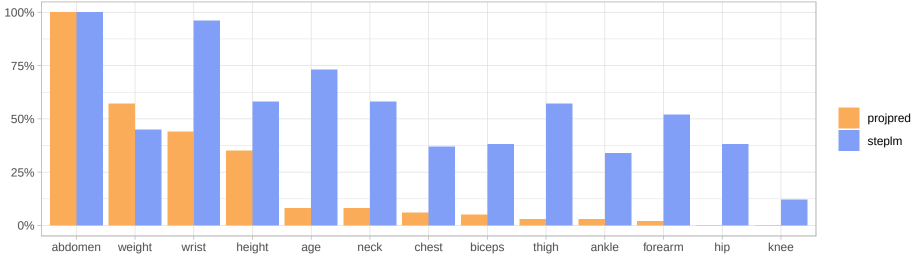
Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets




- Reduced variability, but in case of noisy finite data, there will be some variability under data perturbation
- projpred uses
  - Bayesian inference for the reference
  - The reference model
  - Projection for submodel inference

# Variability under data perturbation

Comparing projection predictive variable selection (projpred) and stepwise maximum likelihood over bootstrapped datasets




- Reduced variability, but in case of noisy finite data, there will be some variability under data perturbation
- projpred uses
  - Bayesian inference for the reference
  - The reference model
  - Projection for submodel inference

# Multilevel regerssion and GAMMs

- `projpred` supports also hierarchical models in `brms`
  Catalina, Bürkner, and Vehtari (2022). Projection predictive inference for generalized linear and additive multilevel models. *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR* 151:4446–4461. https://proceedings.mlr.press/v151/catalina22a.html

# Scaling

- So far the biggest number of variables we've tested is 22K
  - 96s for creating a reference model
  - 14s for projection predictive variable selection

# Intro paper and brms and rstanarm + projpred examples

- McLatchie, Rögnvaldsson, Weber, and Aki Vehtari (2023). Robust and efficient projection predictive inference. https://arxiv.org/abs/2306.15581
- https://mc-stan.org/projpred/articles/projpred.html
- https://users.aalto.fi/~ave/casestudies.html
- Fast and often sufficient if $n \gg p$

```
varsel <- cv_varsel(fit, method='forward', cv_method='loo',
validate_search=FALSE)
```

- Slower but needed if not $n \gg p$

```
varsel <- cv_varsel(fit, method='forward', cv_method='kfold,
K=10, validate_search=TRUE)
```

- If $p$ is very big

```
varsel <- cv_varsel(fit, method='L1, cv_method='kfold, K=5,
validate_search=TRUE)
```