Model Selection

Phil Boonstra, Univeristy of Michigan Biostatistics 699,
Winter 2023

https://github.com/psboonstra/ModelSelection699

Does this model spark joy?

Definition model selection is "estimating the performance of different models in order to choose the best one" (Hastie, Tibshirani, and Friedman 2009)

What it might include

- Choosing structure (GLM, nonlinear model, generalized additive model, discrete/continuous time model, non-likelihood-based [tree, other machine learning methods])
- What variables to include
- How to parametrize (if necessary)

Theory versus practice

Process in BIOS 650, 651, 653, 675, etc is usually

1. fit model
2. report results

Process in many data analyses is

1. explore data
2. make preliminary decisions
3. decide between models
4. fit final model
5. report results
6. assess performance

This lecture talks about some ideas related to step 3. And gives some implications on their impact of steps 5 and 6

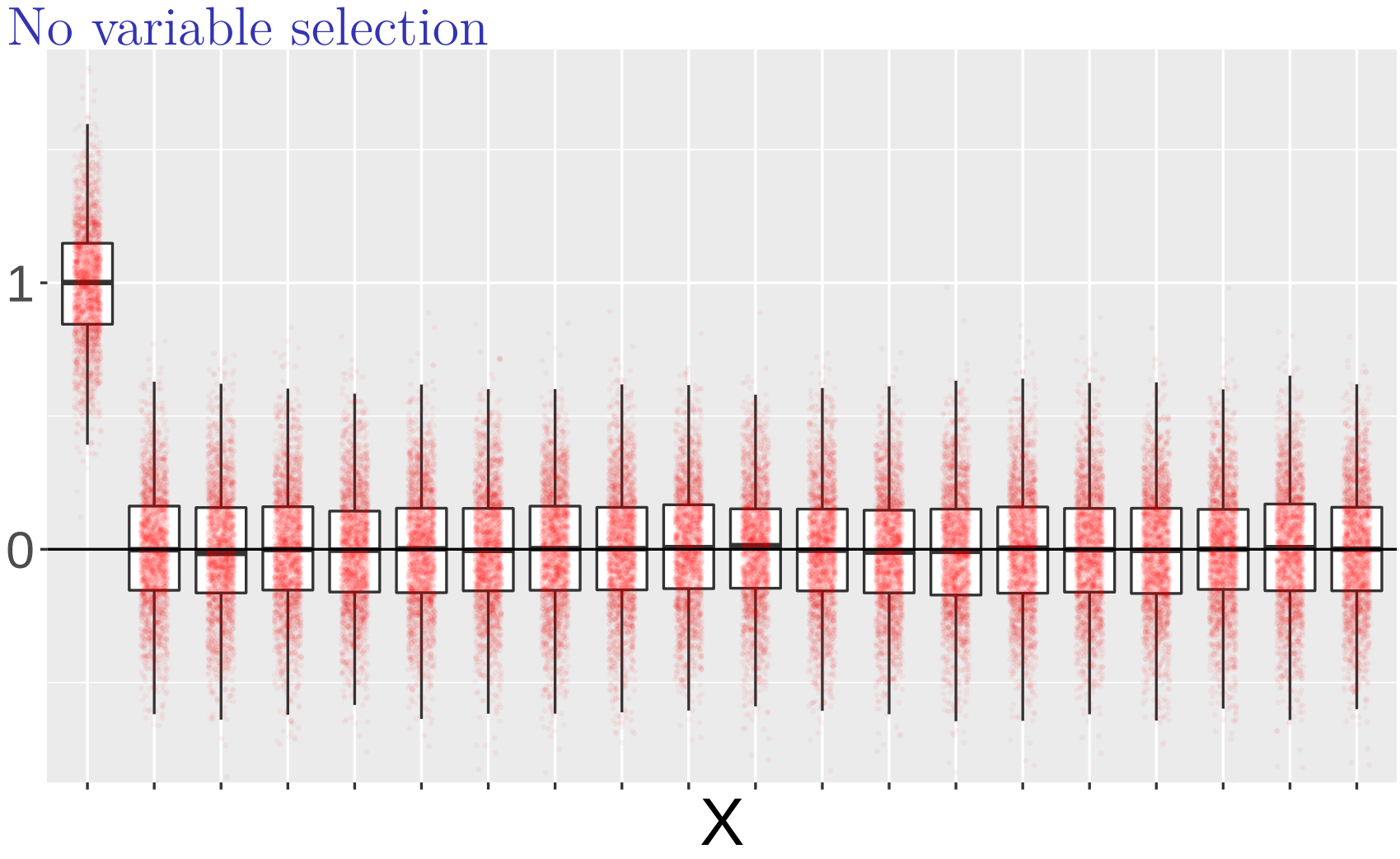
Simulation study to fix ideas

- $X \overset{iid}{\sim} MVN(0, \Sigma(\rho))$; $\epsilon \overset{iid}{\sim} N(0,1)$
- $Y = X\beta + \sigma\epsilon$
- length of β is p
- n observations from model
- $R^2 = \text{var}(X\beta)/[\text{var}(X\beta) + \sigma^2]$
- Compare two approaches:
- No variable selection: fit multiple regression with all variables
- Variable selection: exclude all variables with univariable p-value > 0.10, then fit multiple regression with selected variables

Simulation study, cont'd

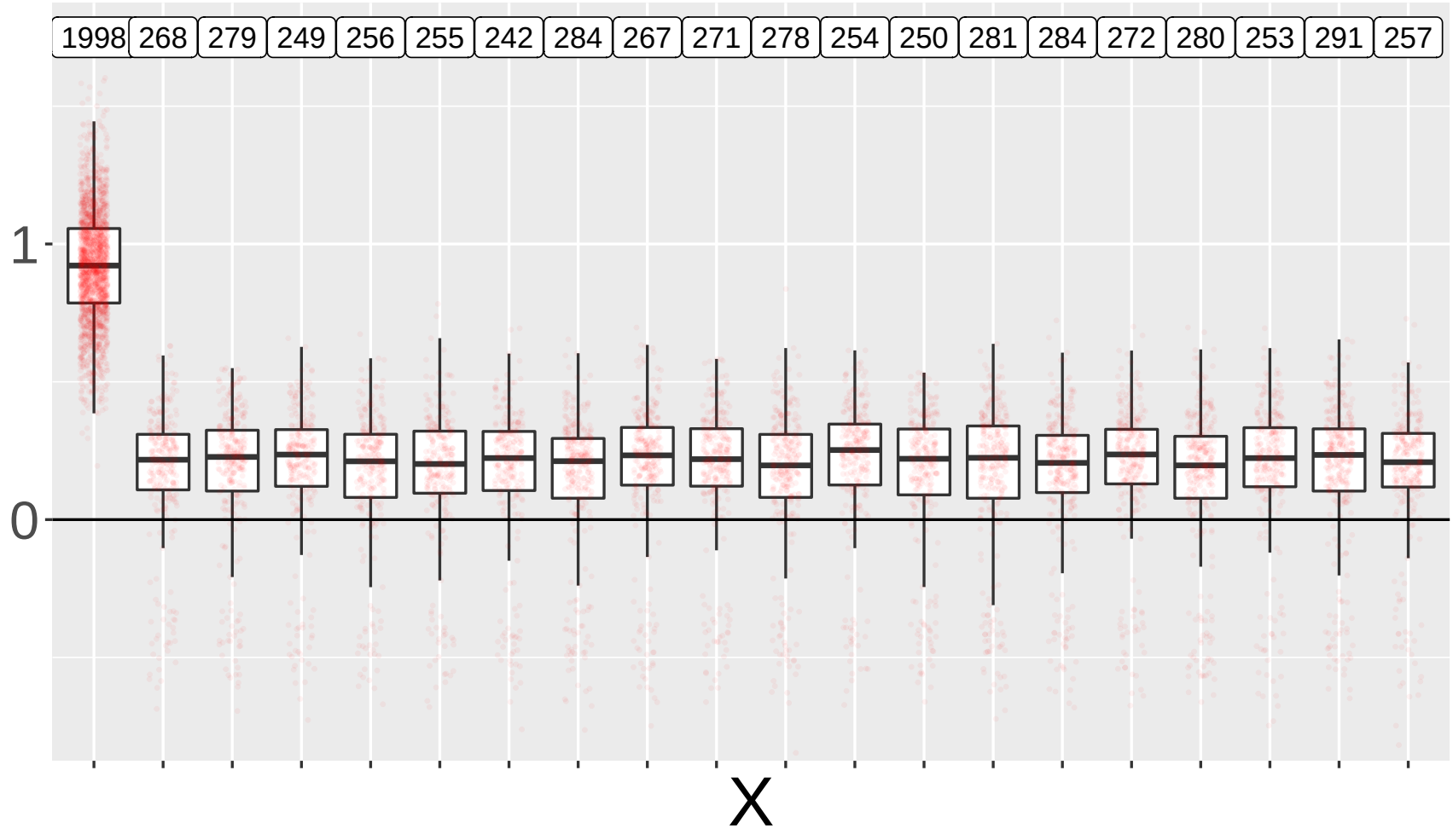
```
library(tidyverse); library(broom);
set.seed(1);
n_sim <- 2e3;
#subjects / simulations
n_subj <- 100;
#compound symmetric correlation
rho <- 0.10;
#true generating betas
p_null <- 19;
#one non-zero beta;
beta <- c(1, numeric(p_null));
#true_rsq implies value of sigma
true_rsq <- 0.2;
#remove all estimates with univariable p-value exceeding
p.value_threshold <- 0.10;
source("stepAICc.R"); # discussed later
source("varselect_sim.R");
```

Distribution of regression estimates



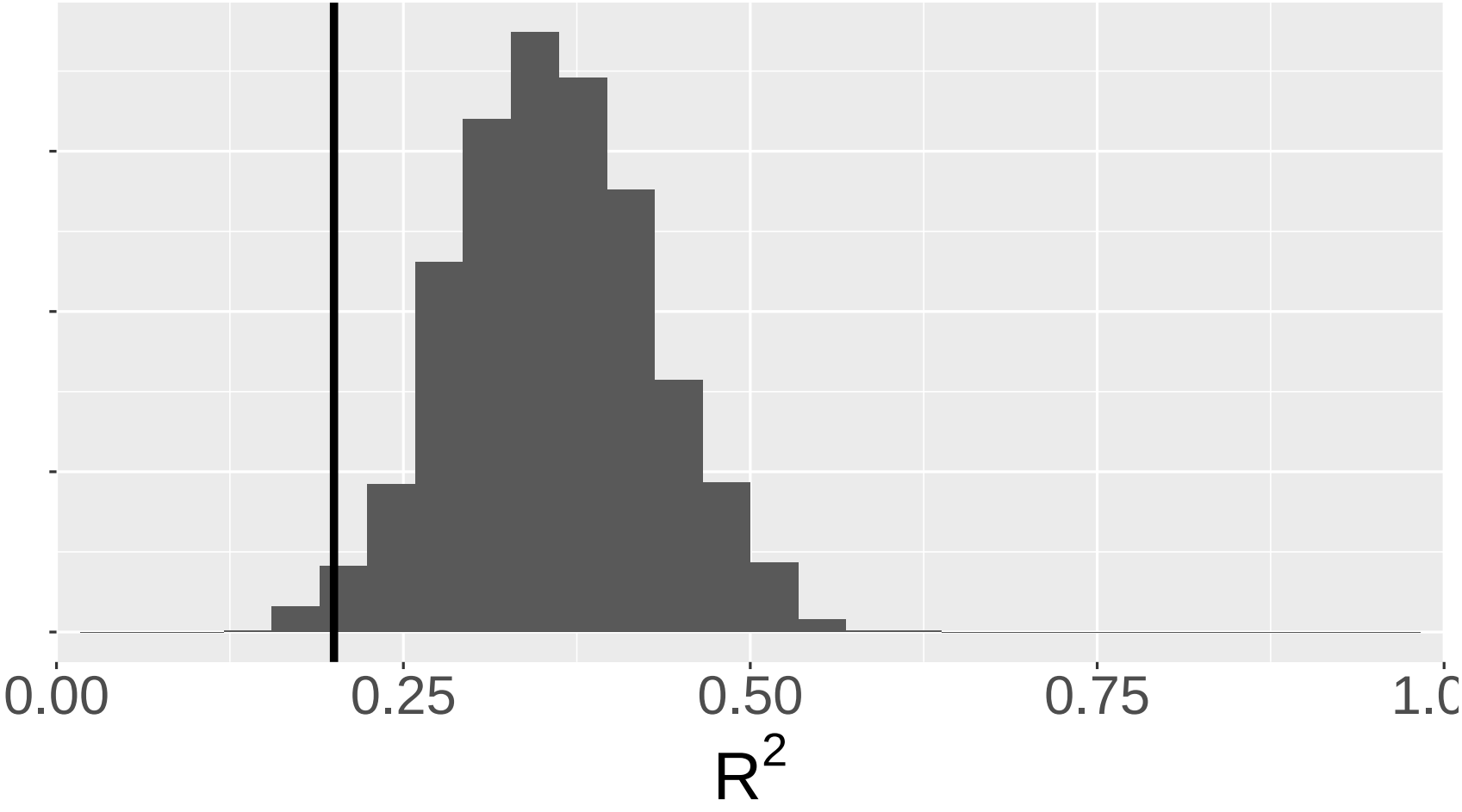
Distribution of regression estimates

Variable selection



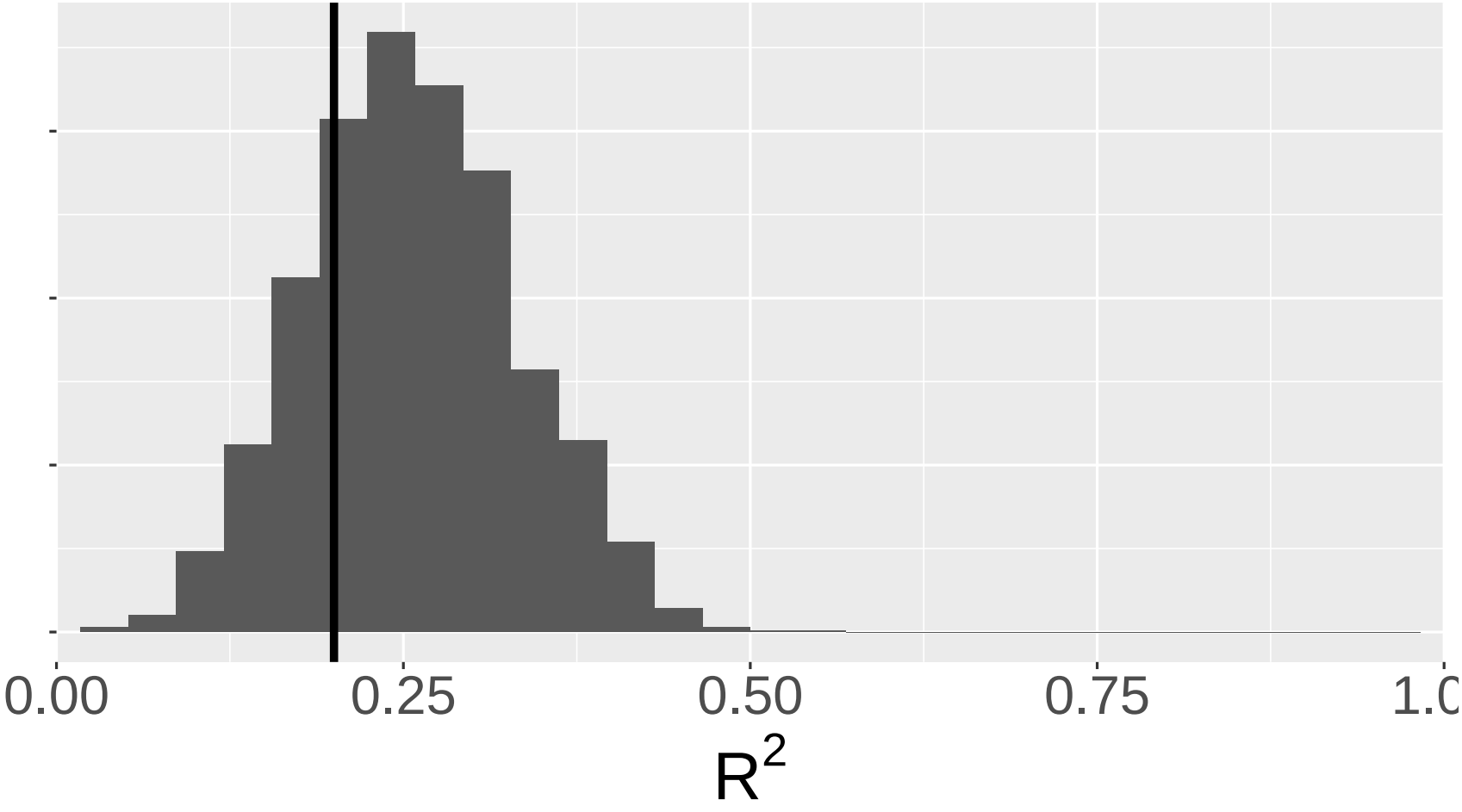
Estimated R^2

No variable selection



Estimated R^2

Variable selection



Comparison

- *No variable selection* approach unbiasedly estimated regression coefficients but may be overfitting data
- *Variable selection* approach biasedly estimated regression coefficients but with lower variance

Careless modeling

Some consequences:

- biased coefficient estimates
- non-nominal significance
- false positive findings
- false certainty
- tendency to overstate proportion of variance explained

Variable selection entails a 'bias-variance' trade-off

What can be done

1. Scientific context should be incorporated wherever possible:

- what is my primary question?
- do prior models exist? how can I incorporate that information?
- are there known constraints? e.g. design features, monotonicity, positivity

2. Data can be used to suggest models

- When used for hypothesis generating, called 'exploratory data analysis' (Tukey 1980)
- SAS / R / STATA have automatic, data-dependent selection routines built in
- When conducted uncritically, 'data dredging' or 'data torture' are better descriptors

Topics we'll be covering

This lecture:

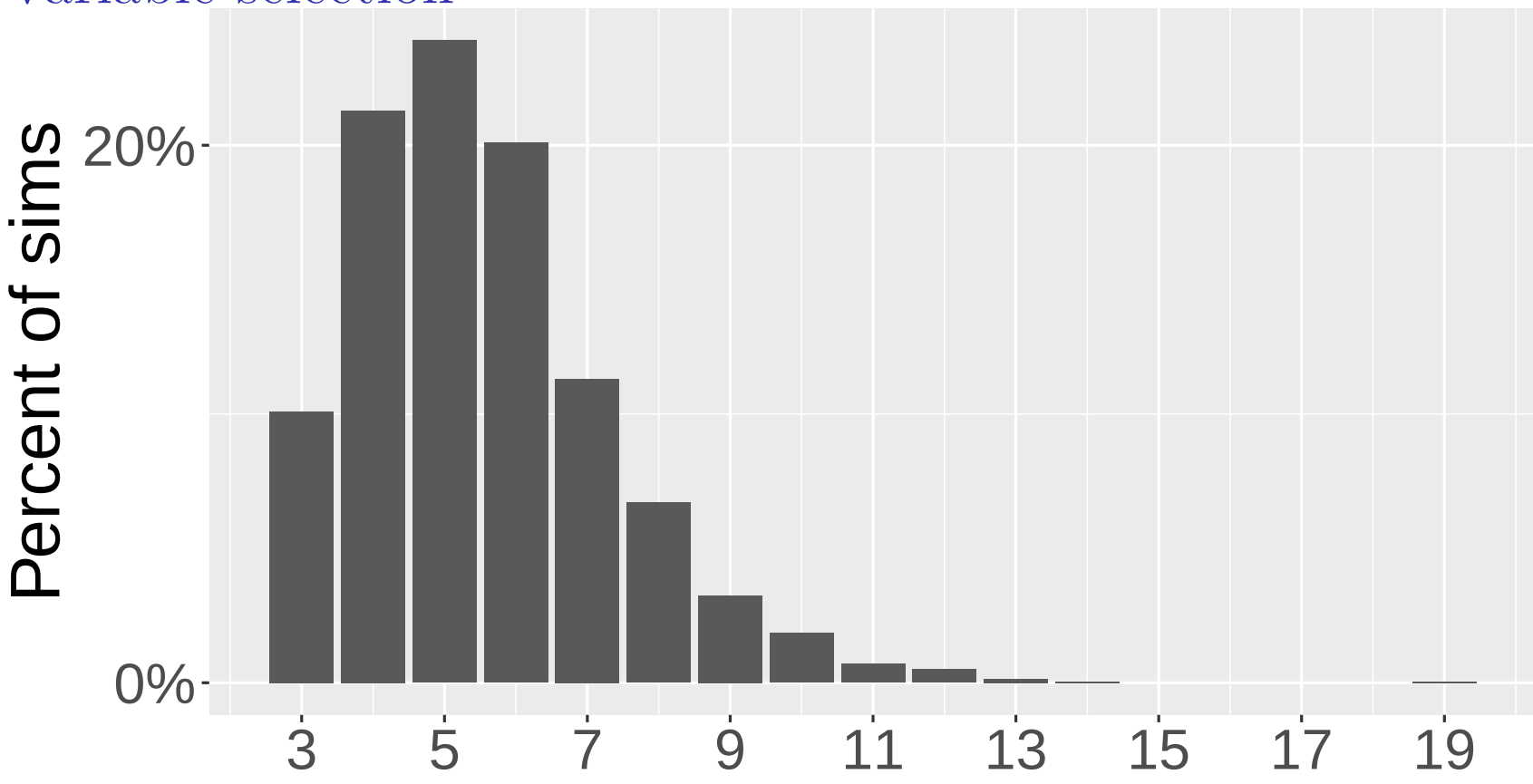
- AIC
- forward / backward selection

Future lecture(s):

- Cross-validation
- LASSO

Model size (including β_0, σ)

Variable selection



Maximized likelihood biased upward relative to expectation

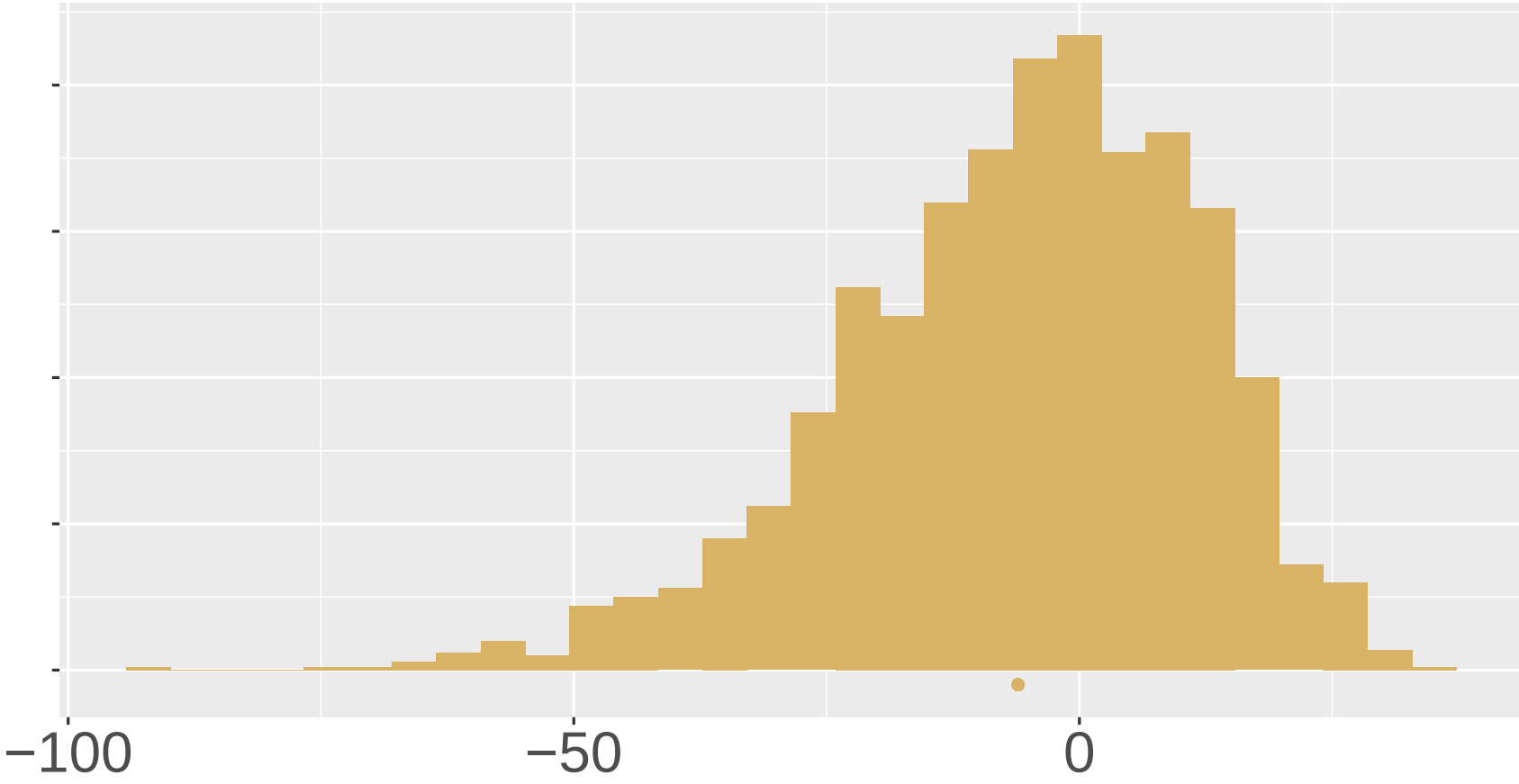
- Ideal setting: **training data** to fit model and independent **validation data** to select model.
- Suppose we have independent observations of outcome given same covariates: $Y_{\text{new}} = X\beta + \sigma\epsilon_{\text{new}}$
- Akaike (1973) asymptotically linked expected log-likelihood of Y_{new} to observed likelihood of Y
- Let $\{\hat{\beta}_0, \hat{\beta}, \hat{\sigma}\}$ be estimated parameters after model building. Then

$$\text{LogLik}_{\text{new}} \approx \text{Loglik} - \text{Model Size}$$

$$\Rightarrow f(Y_{\text{new}}|\hat{\beta}_0, \hat{\beta}, \hat{\sigma}) \approx f(Y|\hat{\beta}_0, \hat{\beta}, \hat{\sigma}) - \#\{\hat{\beta} \neq 0\} - 2$$

$\text{LogLik}_{\text{new}} - \text{Loglik} + \text{Model Size}$

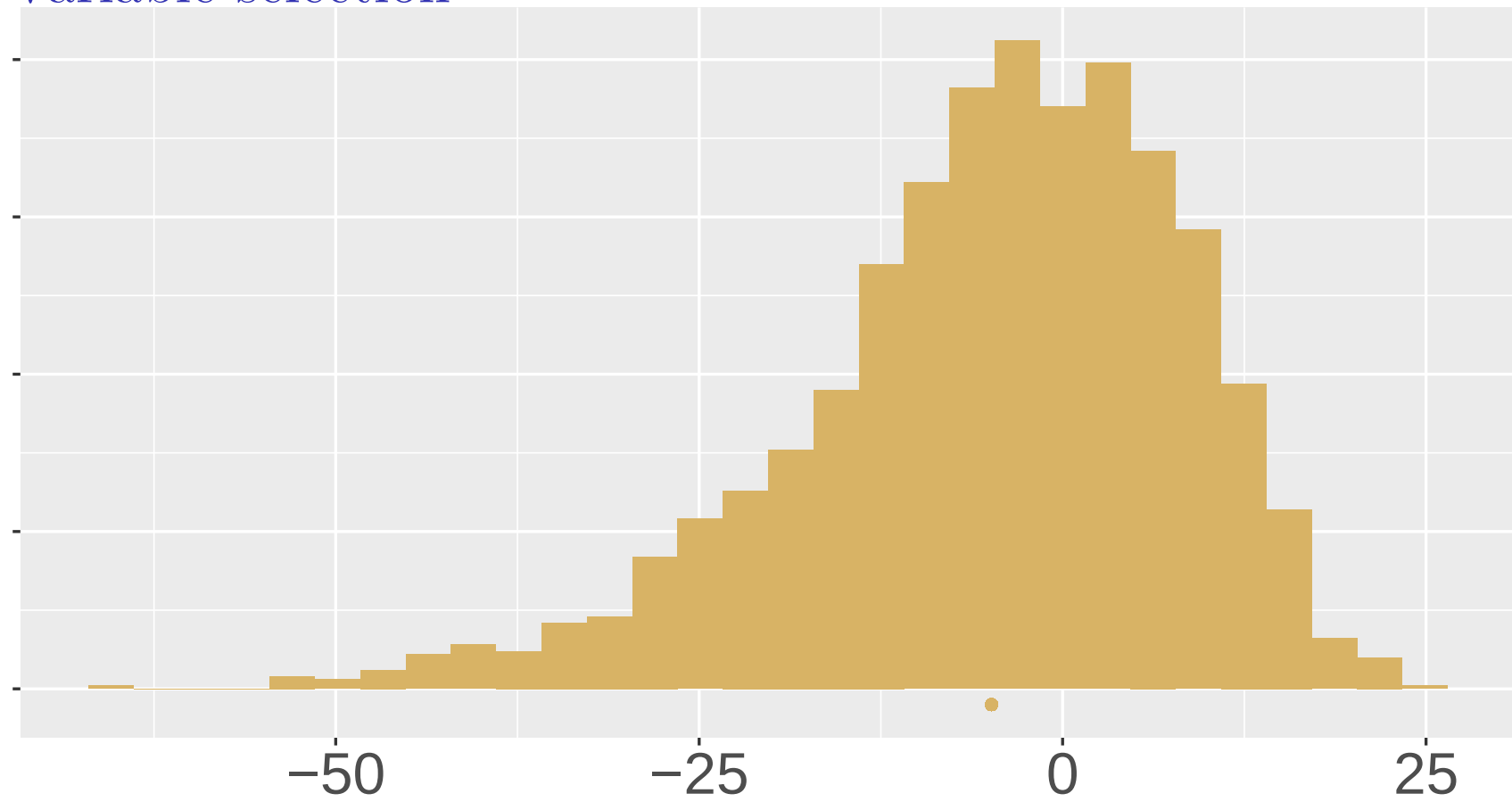
No variable selection



(small dot indicates mean value)

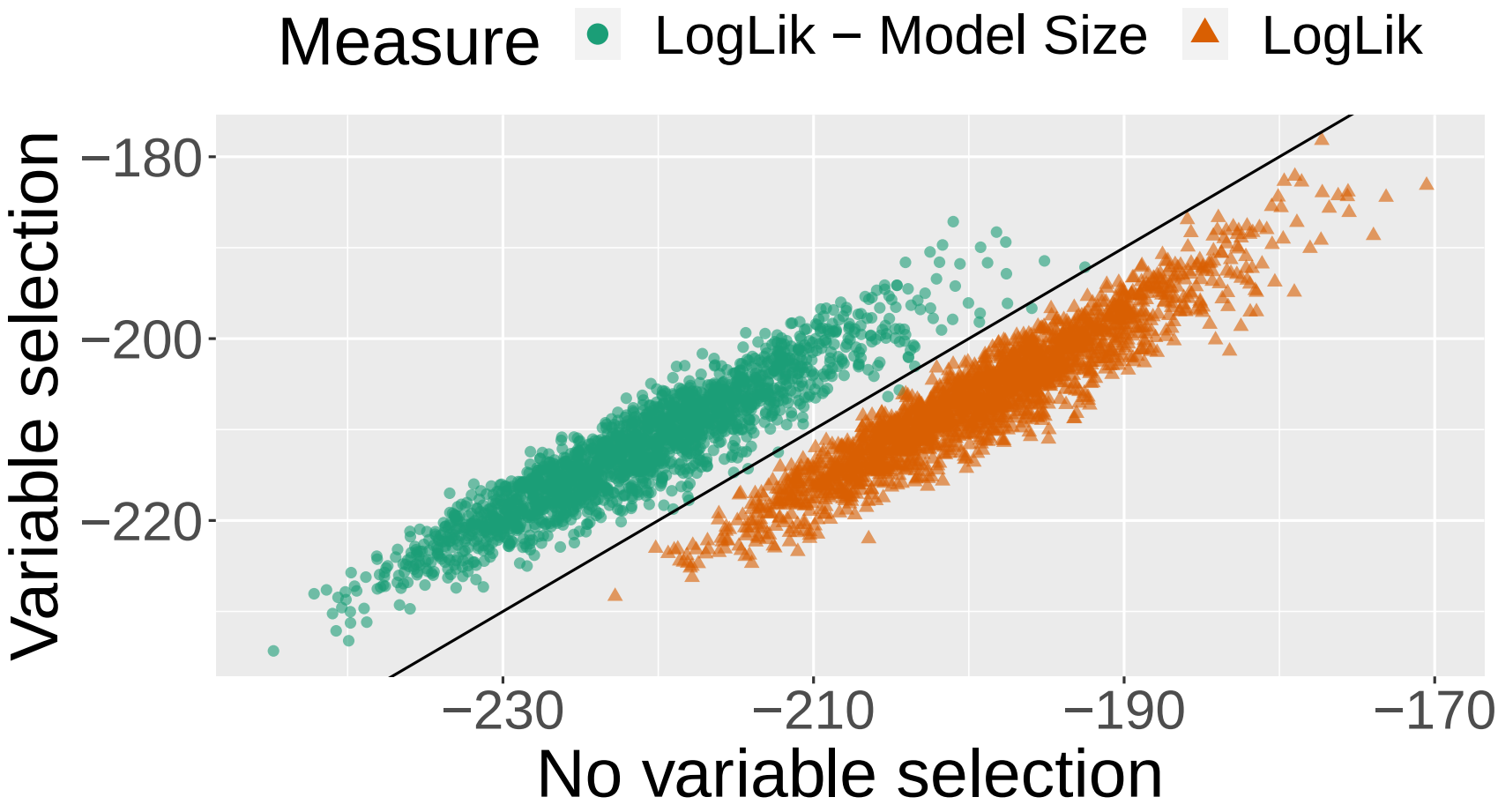
$\text{LogLik}_{\text{new}} - \text{Loglik} + \text{Model Size}$

Variable selection



(small dot indicates mean value)

Comparison of observed vs. adjusted likelihoods



Akaike's Information Criterion (AIC)

- **AIC** defined by $-2 \times$ Loglik $+ 2 \times$ Model Size (-2 multiplier being for "historical reasons")

It is...

- metric useful for comparing relative distance of various fitted models to true generating model (smaller is better)
- approximation of expected log-likelihood of your fitted model when sample size is large relative to number of predictors
- more than intuition that subtracting penalty from model fit is a reasonable thing to do

Akaike's Information Criterion (AIC)

- **AIC** defined by $-2 \times \text{Loglik} + 2 \times \text{Model Size}$ (-2 multiplier being for "historical reasons")

It is *not*...

- absolute measure of model fit
- good approximation of its estimand when sample size is small relative to number of predictors
- useful to compare models fit to non-identical sets of observations
- useful to compare models for different transformations of outcome

AIC and REML

- In general, AIC requires likelihoods given ML-estimated parameters, not REML-estimated (see Section 4.7.2, Gałecki and Burzykowski (2013))
- Possible to compare models using ML-estimates, then report REML-estimates from selected model
- See also https://tinyurl.com/yagpx4bv (StackExchange:CrossValidated)

Small-sample AIC

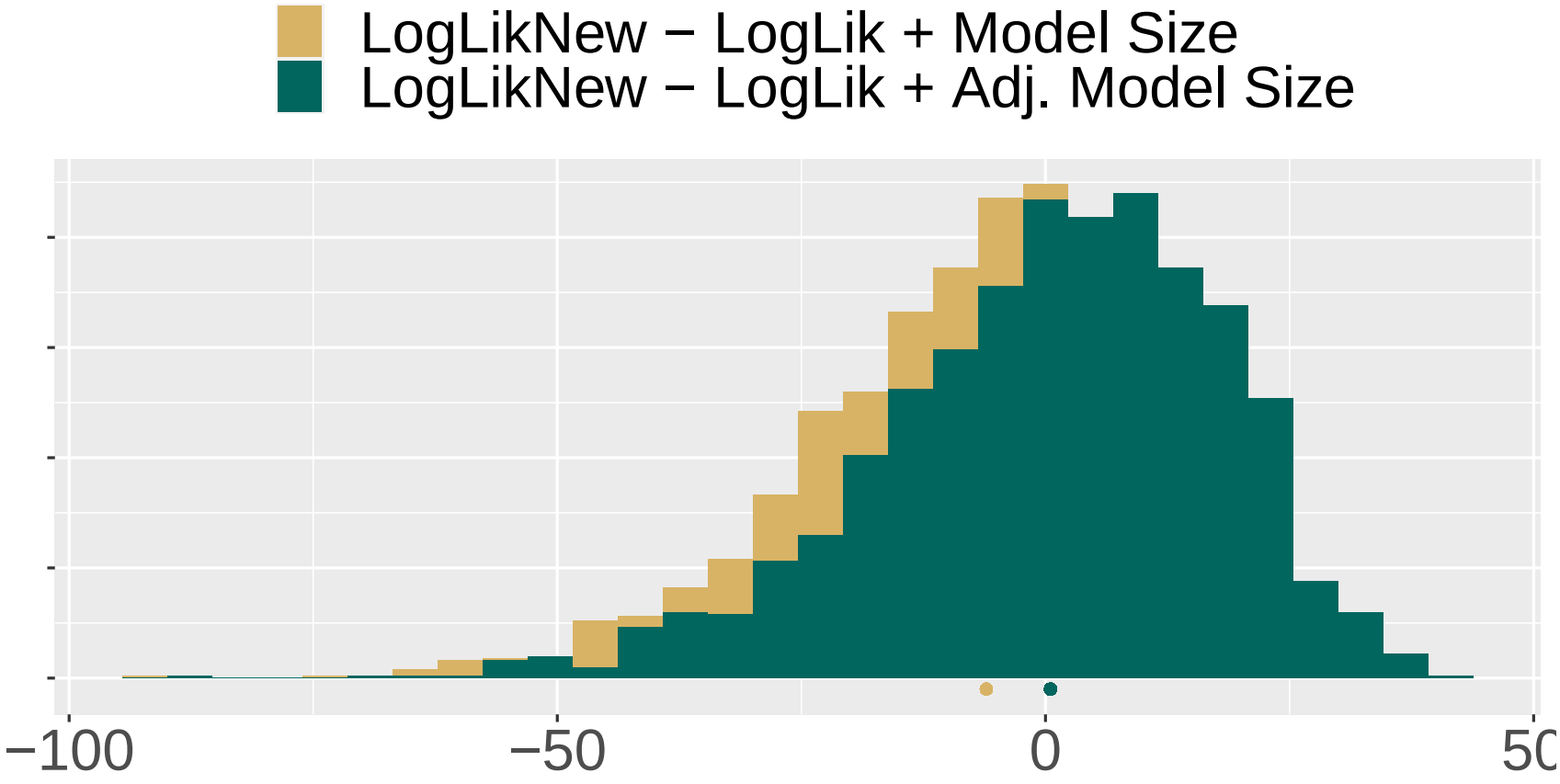
- Hurvich and Tsai (1989) proposed a faster-converging "corrected" AIC with an adjusted estimate of model size:

$$\text{AIC}_C = -2 \times \text{Loglik} + 2 \times \text{Model Size} \times \left(\frac{n}{n - \text{Model Size} - 1} \right)$$

- To account for $p \gg n$, Boonstra, Mukherjee, and Taylor (2015) suggest instead positive part of correction:
$\left(\frac{n}{n - \text{Model Size} - 1} \right)_+$
- Phil's opinion: always use AIC_C

$$\text{LogLik}_{\text{new}} - \text{Loglik} + \text{Model Size} \times \left(\frac{n}{n - \text{Model Size} - 1}\right)$$

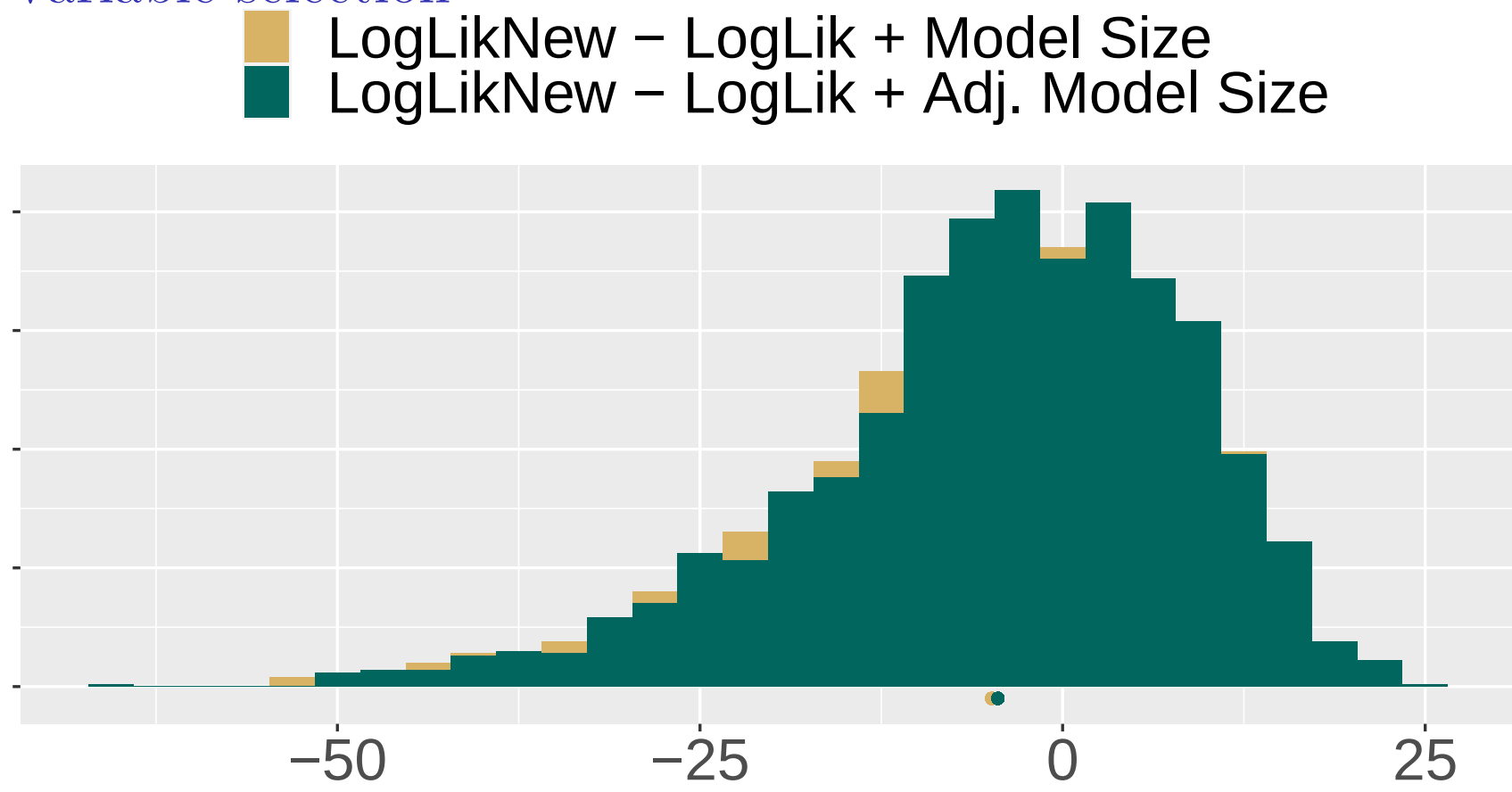
No variable selection



(small dots indicate mean values)

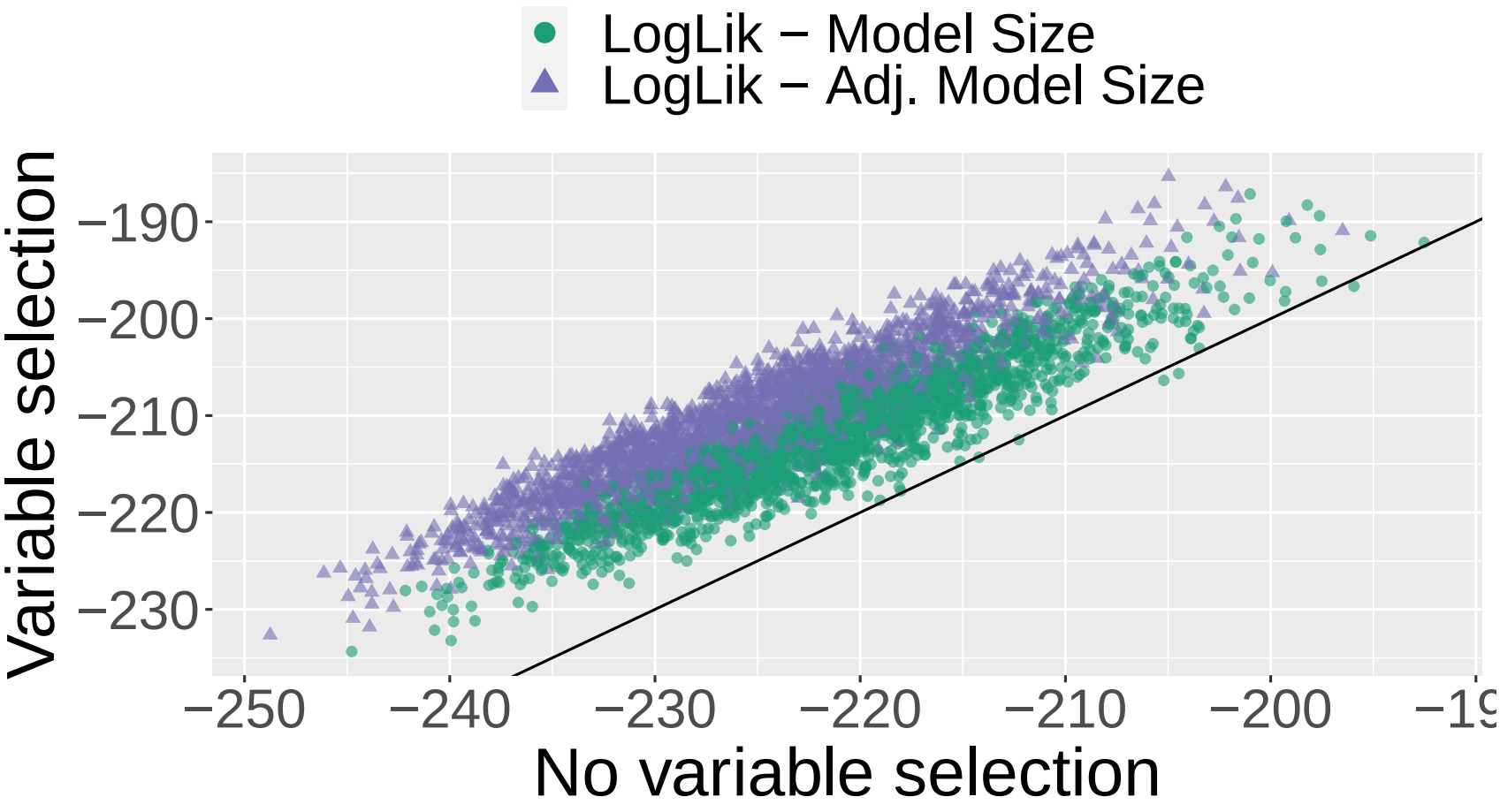
$$\text{LogLik}_{\text{new}} - \text{Loglik} + \text{Model Size} \times \left(\frac{n}{n - \text{Model Size} - 1}\right)$$

Variable selection



(small dots indicate mean values)

AIC_C favors smaller models



Selecting variables

- When considering what variables from a list to include, AIC / AIC_C are means of comparing them
- But cannot compare all 2^p possible models when p exceeds 20–30.

This grief has a gravity, it pulls me down But a tiny voice whispers in my mind You are lost, hope is gone But you must go on... Just do the next right thing Take a step, step again It is all that I can to do The next right thing

This grief has a gravity, it pulls me down But a tiny voice whispers in my mind You are lost, hope is gone But you must go on... Just do the next right thing Take a step, step again It is all that I can to do The next right thing

Anna (Frozen 2)

Sequential selection procedures

- Forward selection starts with intercept, adds variables sequentially that "improve" model by greatest amount
- Backward selection starts with full model, subtracts variables presence of which "worsen" model by greatest amount

`MASS::stepAIC()` function does automatic forward/backward selection using AIC as measure of model fit

AIC is defined only up to constant

R functions `AIC()` and `extractAIC()` (which `stepAIC()` uses) employ different constants. From `extractAIC` help page:

> *For linear models with unknown scale (i.e., for lm and aov), -2 log L is computed from the deviance and uses a different additive constant to logLik and hence AIC.*

For those interested, the difference between `AIC()` and `extractAIC()` for linear models is $n(\log(2\pi) + 1)$

Example: Breast Cancer Diagnosis data

Digitized image of fine needle aspirate (FNA) of breast mass from 569 patients

Street, Wolberg, and Mangasarian (1993) Mangasarian, Street, and Wolberg (1995)

Outcome Clinical diagnosis (malignant or benign)

Predictors

a. radius (mean of distances from center to points on the perimeter)
b. texture (standard deviation of gray-scale values)
c. perimeter
d. area
e. smoothness (local variation in radius lengths)
f. compactness (perimeter^2 / area - 1.0)
g. concavity (severity of concave portions of the contour)
h. concave points (number of concave portions of the contour)
i. symmetry
j. fractal dimension ("coastline approximation" - 1)

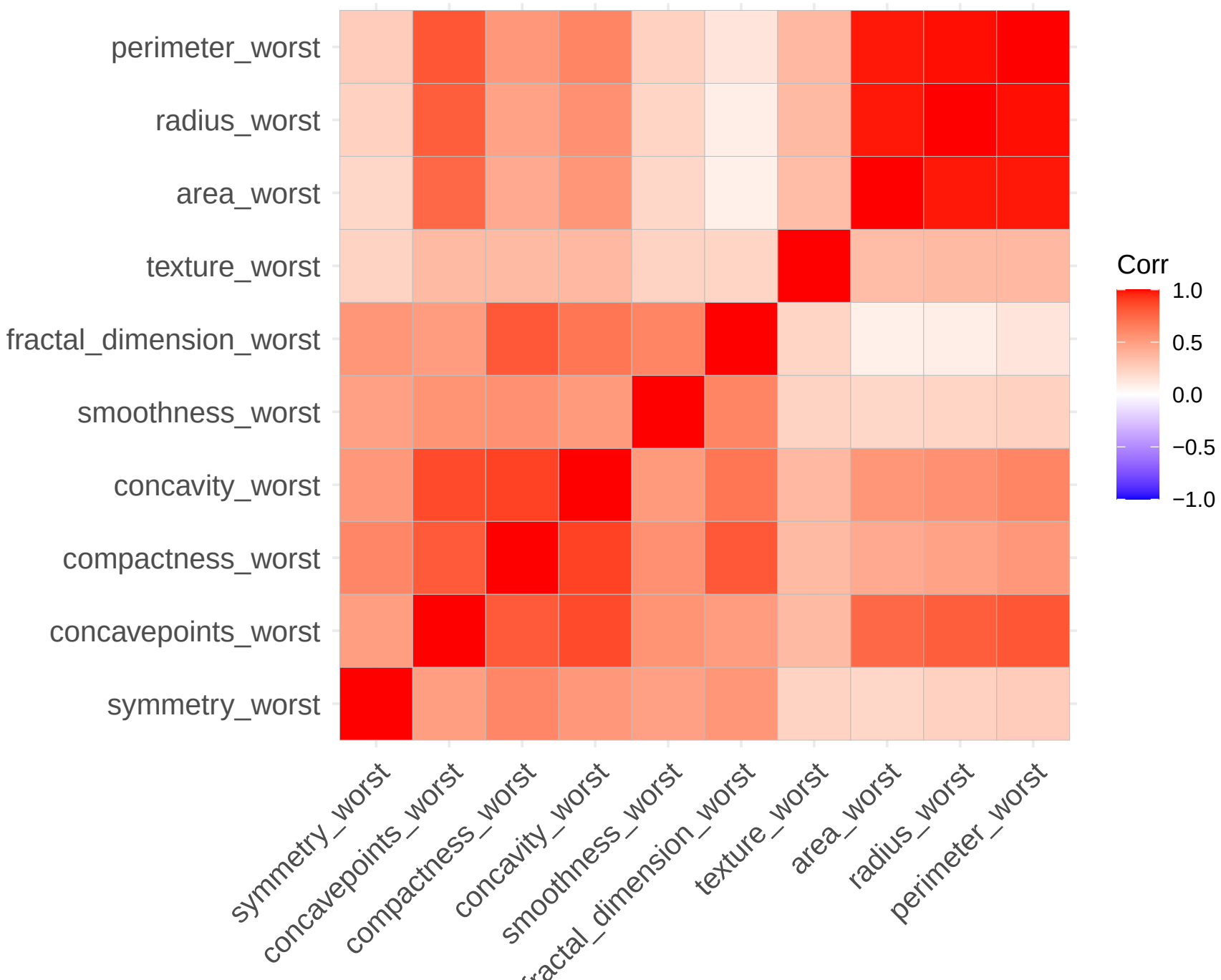
Data include mean, standard error, and worst measurements.

Available at https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

```r
breast_dx <-
  read_csv("bdiag.csv", show_col_types = FALSE) %>%
  # Translate M/D into 1/0
  mutate(malignant = 1 * (diagnosis == "M")) %>%
  # Drop errant space in 'concave points_mean' variable nam
  rename_with(~str_replace(string = ., pattern = " ", repla
  # Focus only on worst measurements
  select(malignant,
         #contains("_mean"),
         #contains("_se"),
         contains("_worst"))
```

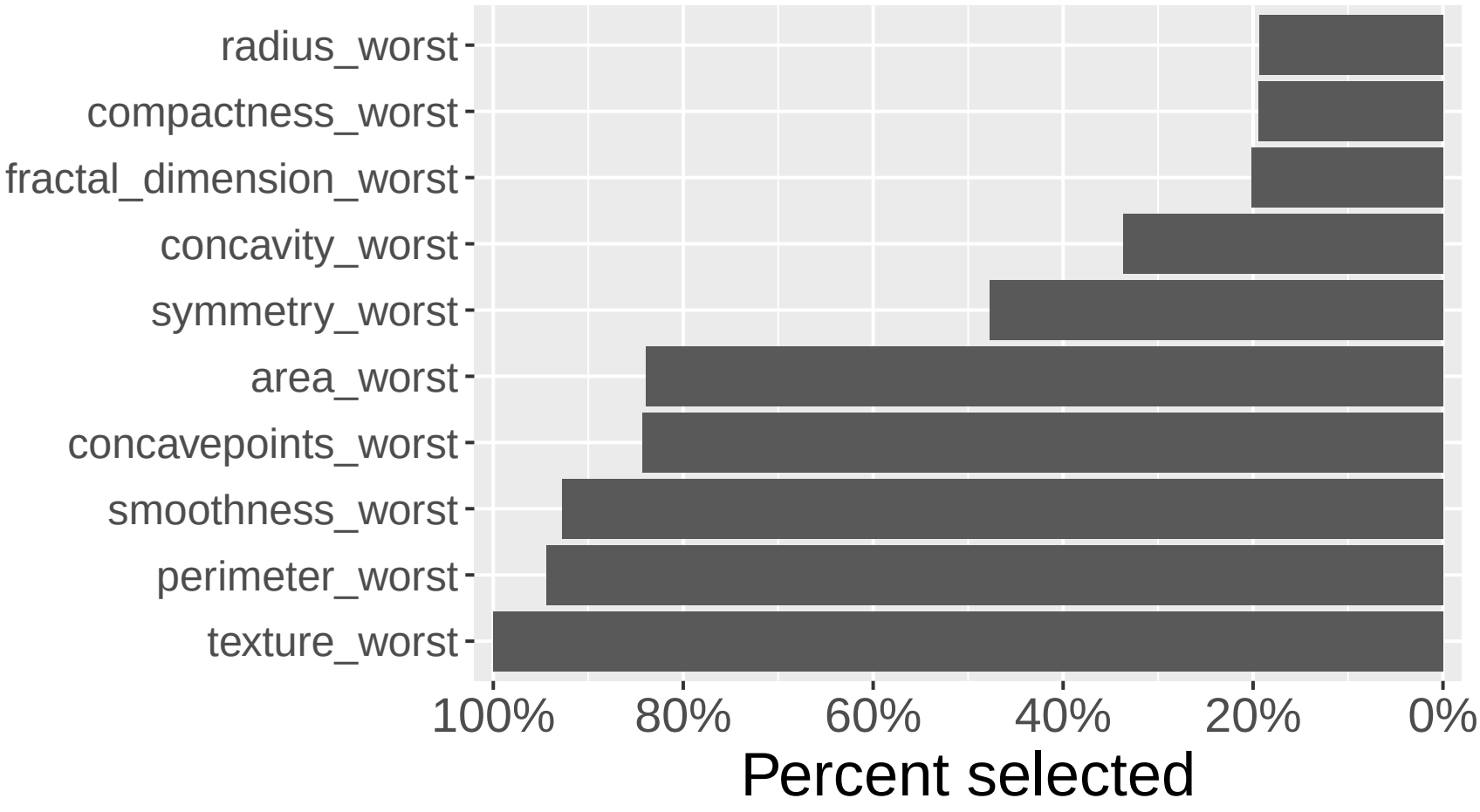
Plan to fit a logistic regression model to predict probability of being malignant/benign using set of *worst* measurements

```
ggcorrplot(cor(select(breast_dx, -malignant)), hc.order = T
```



Bootstrap demonstrates selection instability

1. Sample equal sized dataset *with replacement* from the original dataset
2. Run forward selection process bootstrap dataset, separately for each criterion
3. Record selected model
4. Repeat steps 1–3 $B = 2000$ times


radius_worst
compactness_worst
fractal_dimension_worst
concavity_worst
symmetry_worst
area_worst
concavepoints_worst
smoothness_worst
perimeter_worst
texture_worst
100%
80%
60%
40%
20%
0%
Percent selected

All models with AICc no more than 0.5 away from original selected models. 141 distinct variable combinations selected across B = 2000 bootstrap datasets

Variable set	Pct. Sel. AICc	Delta(Obs. AICc)
area, compactness, concavepoints, concavity, smoothness, symmetry, texture	0.7%	-1.23
area, concavepoints, smoothness, texture	1.2%	-1.20
area, concavepoints, smoothness, symmetry, texture	0.9%	-1.06
area, compactness, concavepoints, smoothness, symmetry, texture	0.2%	-0.85
area, concavepoints, fractal_dimension, smoothness, symmetry, texture	0.2%	-0.27
area, compactness, concavepoints, concavity, smoothness, texture	0.2%	-0.09
area, concavepoints, radius, smoothness, texture	0.3%	-0.04
area, concavepoints, perimeter, smoothness, texture	20.9%	-0.02
area, concavepoints, perimeter, smoothness, symmetry, texture	12.2%	0.00
area, concavepoints, radius, smoothness, symmetry, texture	0.1%	0.03

Comparison of three models fit to observed breast DX data

Variable set	Pct. Sel. AICc	Delta(Obs. AICc)
area, compactness, concavepoints, concavity, smoothness, symmetry, texture	0.7%	-1.23
area, concavepoints, perimeter, smoothness, texture	20.9%	-0.02
area, concavepoints, perimeter, smoothness, symmetry, texture	12.2%	0.00

Comparison of three models fit to observed breast DX data

log-OR (standardized log-OR)

	Forward selected	Most common (bootstrap)	Smallest AICc (bootstrap)
area_worst	0.02(11.6)	0.02(11.1)	0.01(8.2)
compactness_worst			-9.31(-1.5)
concavepoints_worst	38.47(2.5)	44.10(2.9)	36.89(2.4)
concavity_worst			5.06(1.1)
perimeter_worst	-0.10(-3.2)	-0.09(-2.9)	
smoothness_worst	45.81(1.0)	46.37(1.1)	52.74(1.2)
symmetry_worst	6.94(0.4)		9.40(0.6)
texture_worst	0.28(1.7)	0.28(1.7)	0.28(1.7)

standardized log-OR = log-OR * standard deviation of covariate

Summary thoughts

AIC, AIC_C require two ingredients

AIC, AIC_C useful model selection technique depending upon being able to do two things:

- calculate likelihood
- count parameters (not always trivial)

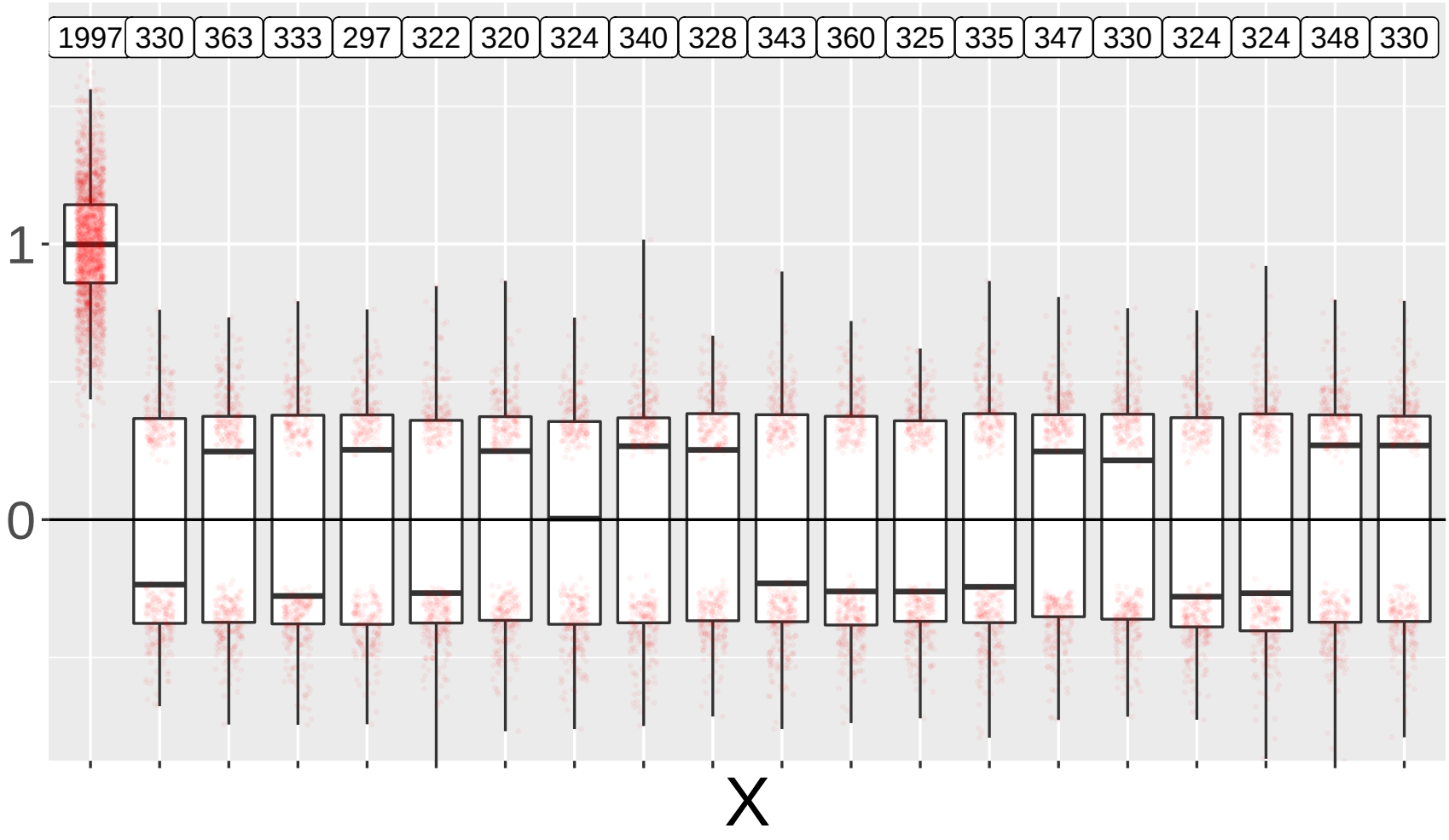
When one or both of these is not calculable, must use other options, e.g. cross-validation / split-sampling

See "Model Selection and Inference" (Burnham and Anderson, 1998) for a nice introduction to theory of AIC (https://link.springer.com/book/10.1007/978-1-4757-2917-7)

AIC and AIC_C do not fix the problems with stepwise selection

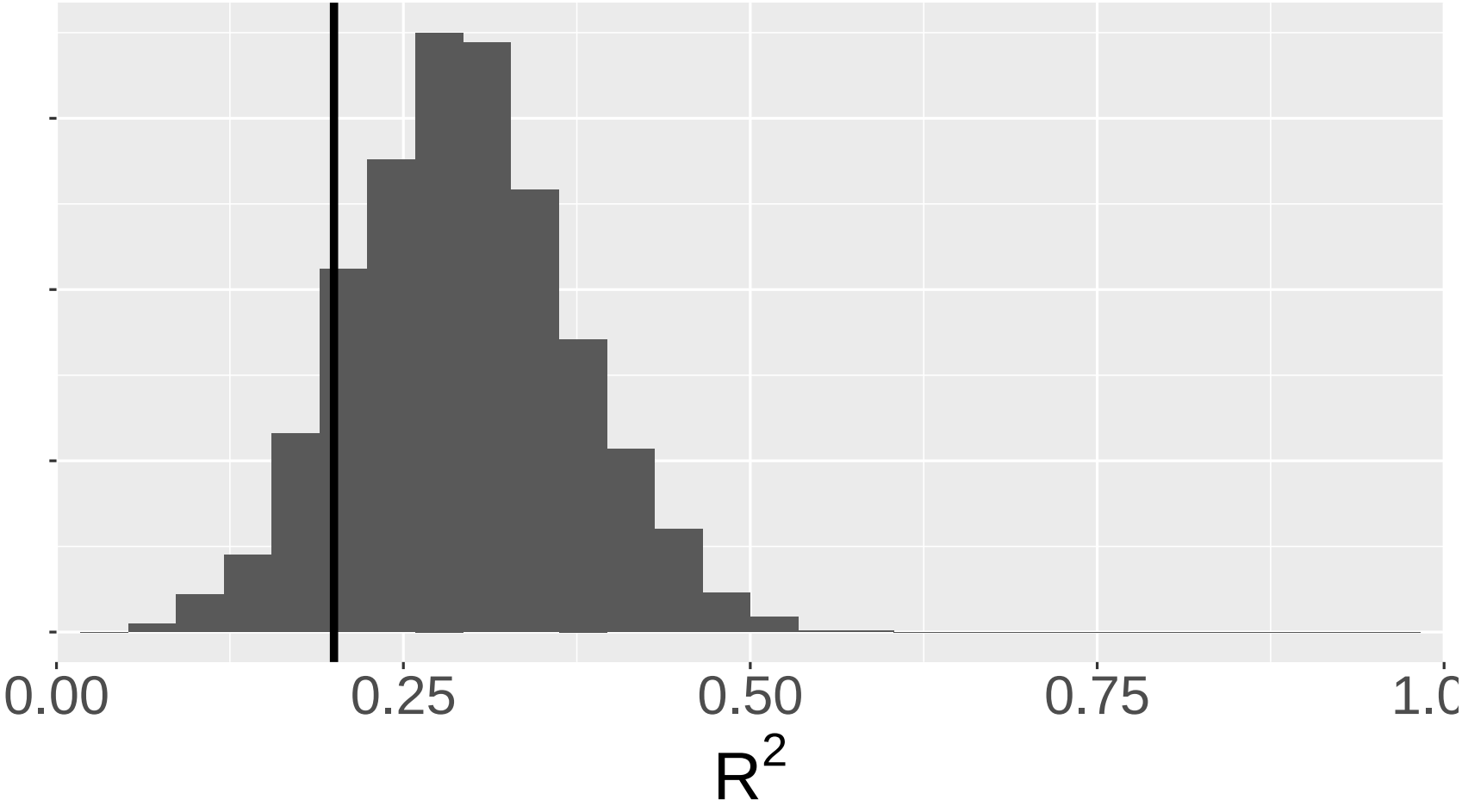
Distribution of regression estimates

AICc+Forward Selection



Estimated R^2

AICc+Forward Selection



There are better options than stepwise selection

- Least absolute shrinkage and selection operator (LASSO)
- Elastic net

Be familiar with your data / interrogate your model

Model selection is not model assessment

- Focus of talk has been on *model selection* ("estimating the performance of different models in order to choose the best one"), distinct from *model assessment* ("having chosen a final model, estimating its prediction error...on new data", Hastie, Tibshirani, and Friedman (2009))
- Standard errors of selected model do not account for selection process. Ideal assessment requires additional (third) data set: *testing data* (future lecture (?))

References

Akaike, Hirotsugu. 1973. "Information Theory and an Extension of the Maximum Likelihood Principle." *Second International Symposium on Information Theory*, 267–81.

Boonstra, Philip S, Bhramar Mukherjee, and Jeremy M G Taylor. 2015. "A Small-Sample Choice of the Tuning Parameter in Ridge Regression." *Statistica Sinica* 25 (3): 1185–1206.

Gałecki, Andrzej, and Tomasz Burzykowski. 2013. *Linear Mixed-Effects Models Using r: A Step-by-Step Approach*. Springer Science & Business Media.

Hastie, T, R Tibshirani, and J Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York, NY: Springer.

Hurvich, Clifford M., and Chih-Ling Tsai. 1989. "Regression and Time Series Model Selection in Small Samples." *Biometrika* 76: 297–307.

Mangasarian, Olvi L, W Nick Street, and William H Wolberg. 1995. "Breast Cancer Diagnosis and Prognosis via Linear Programming." *Operations Research* 43 (4): 570–77.

Street, W Nick, William H Wolberg, and Olvi L Mangasarian. 1993. "Nuclear Feature Extraction for Breast Tumor Diagnosis." In *Biomedical Image Processing and Biomedical Visualization*, 1905:861–70. International Society for Optics; Photonics.

Tukey, John W. 1980. "We Need Both Exploratory and Confirmatory." *The American Statistician* 34 (1): 23–25.