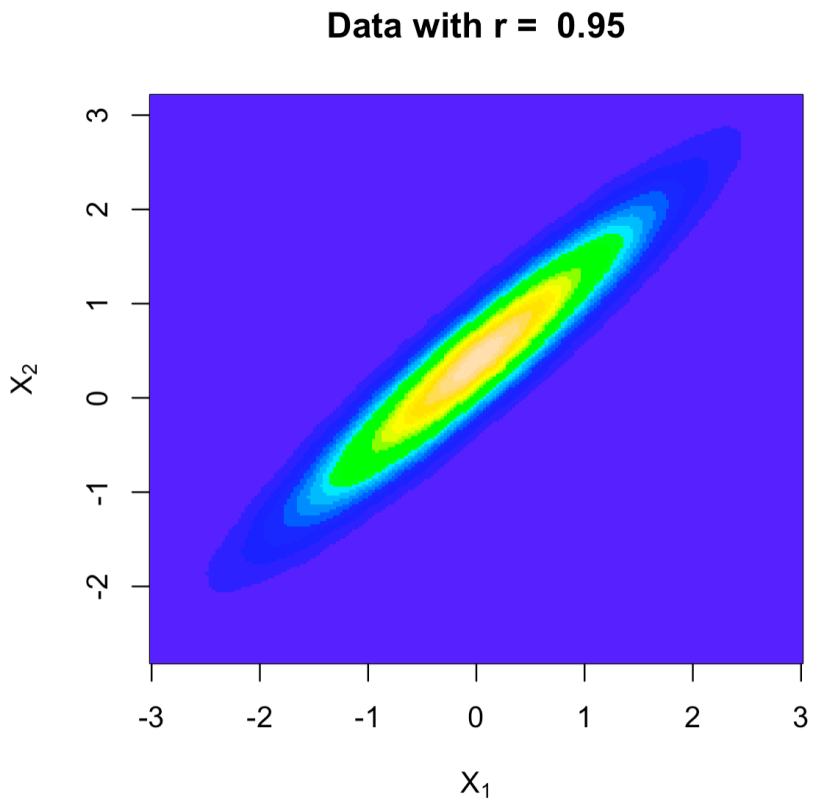
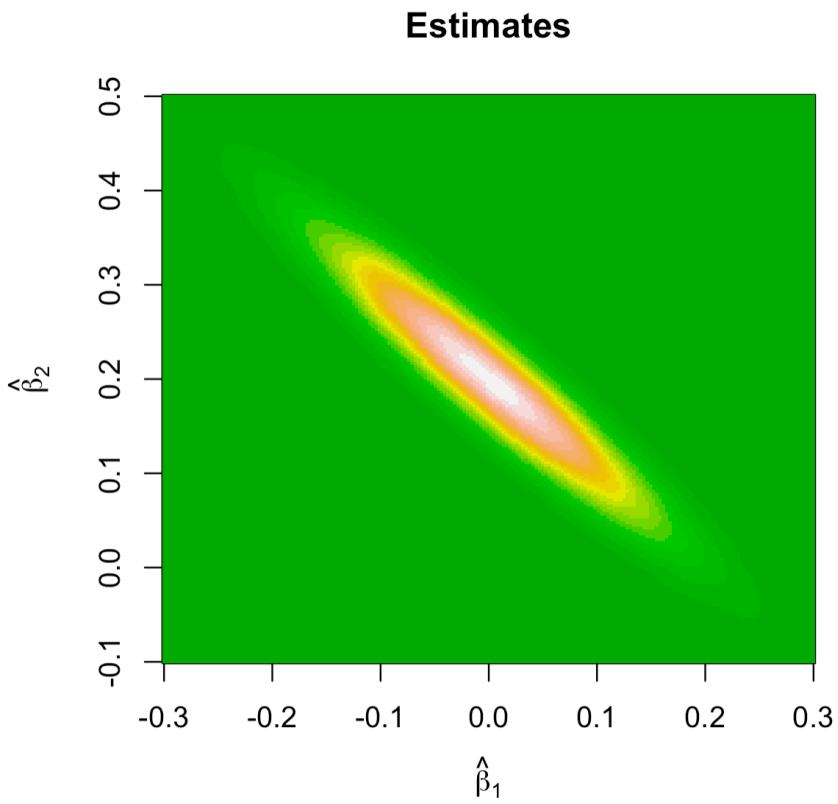
CORRELATED PREDICTORS





- $Y = 0.2\ X_2 + e$,
- $\text{cor}(X_1, X_2) = r$
- How do estimates of coeffs of X_1 and X_2 behave as function of r?

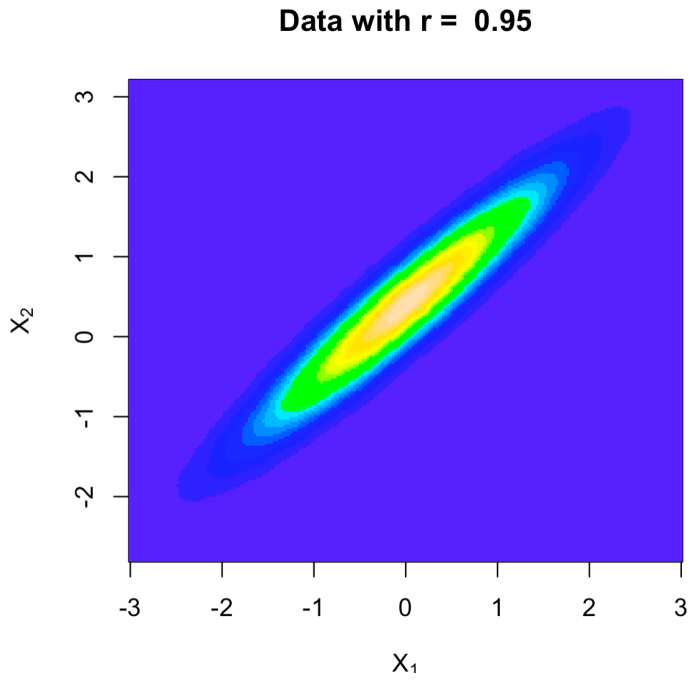
$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$$

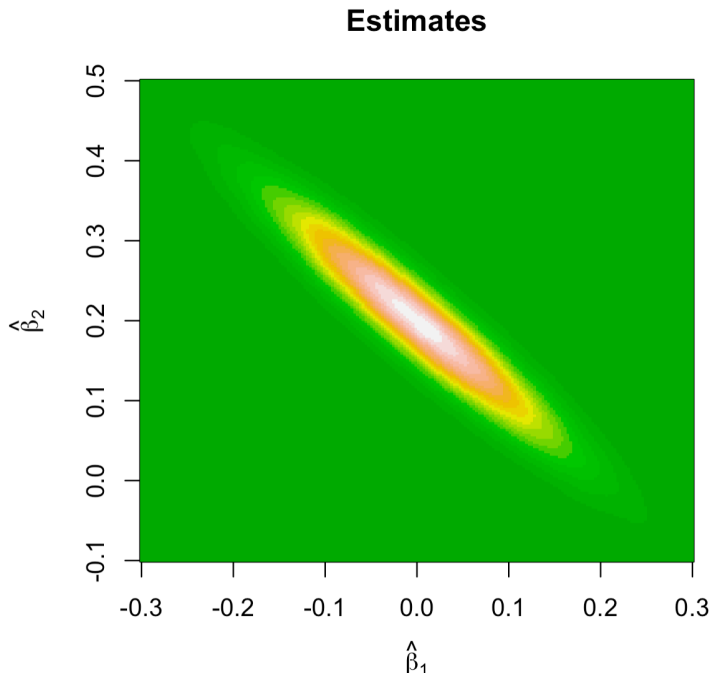
$$\sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} = \frac{\sigma^2}{n} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}^{-1} = \frac{\sigma^2}{n(1-r^2)} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$$

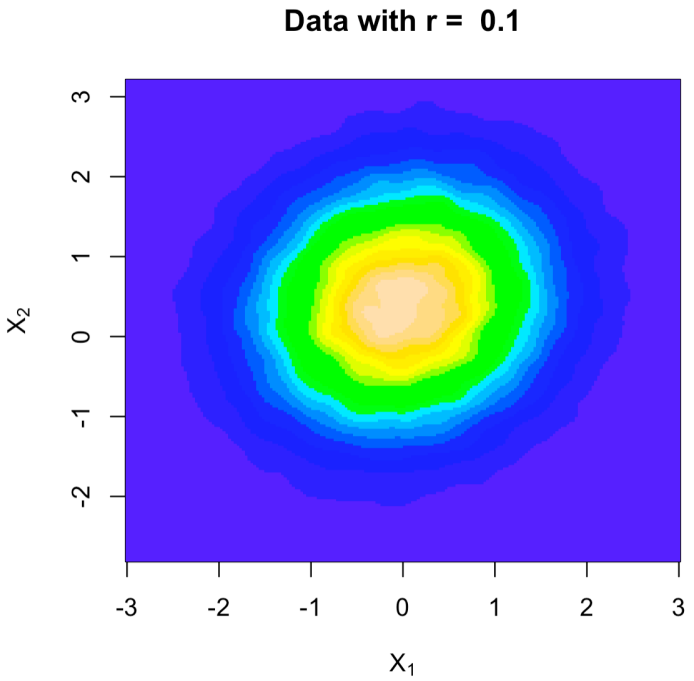
$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}\right)$$

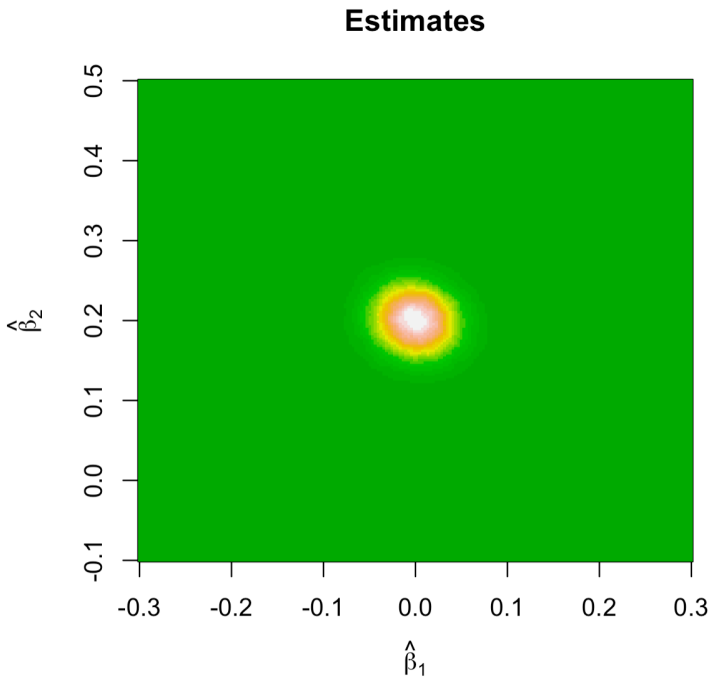
$$\sigma^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1} = \frac{\sigma^2}{n} \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}^{-1}$$

$$= \frac{\sigma^2}{n(1-r^2)} \begin{pmatrix} 1 & -r \\ -r & 1 \end{pmatrix}$$









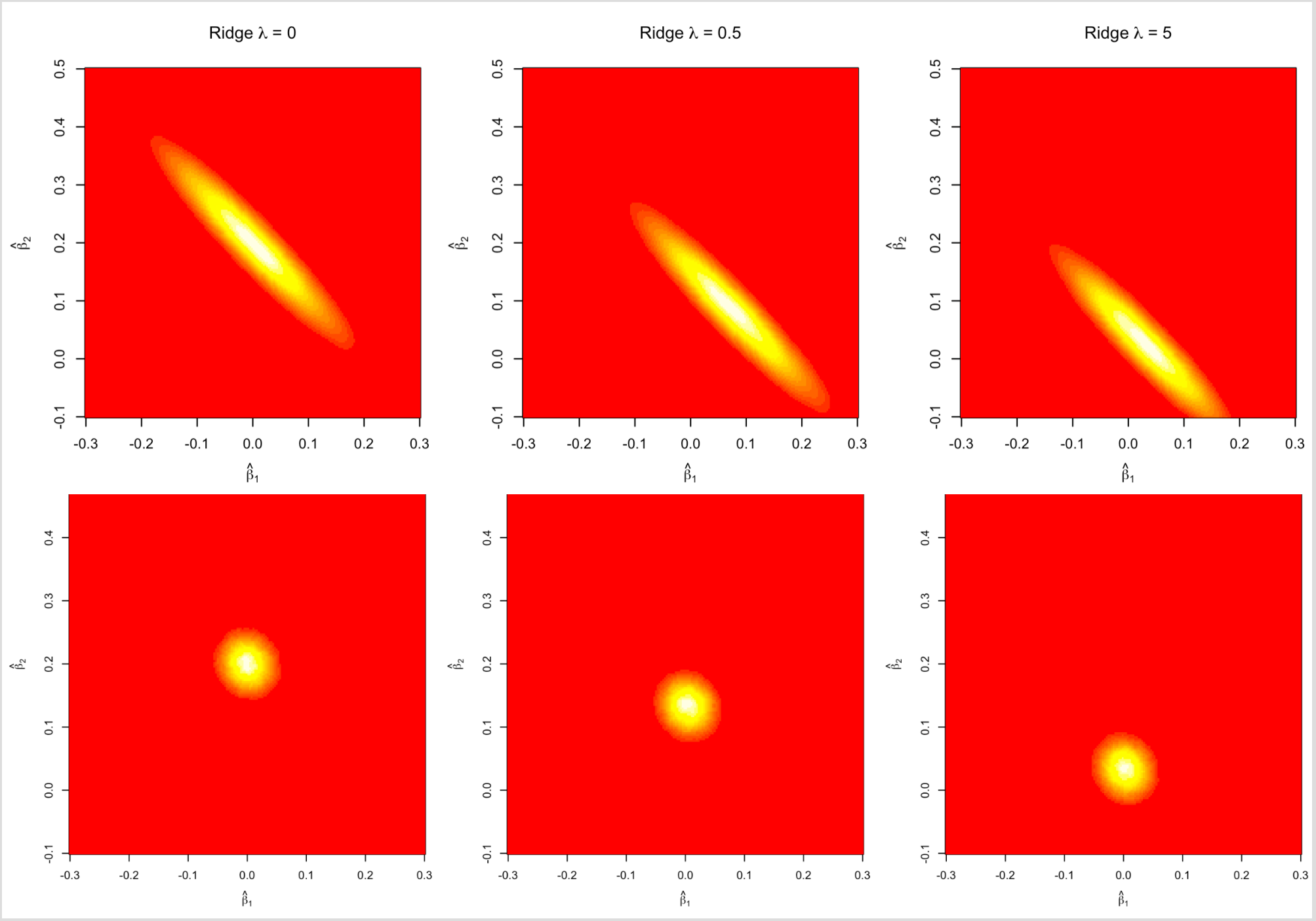
Note how SEs of estimators increase with |r|.

Intuitively this is because with highly correlated variables it is less clear how to split the effects between variables: There are many almost equally likely ways.

RIDGE REGRESSION ESTIMATOR

r = 0.95

r = 0.1

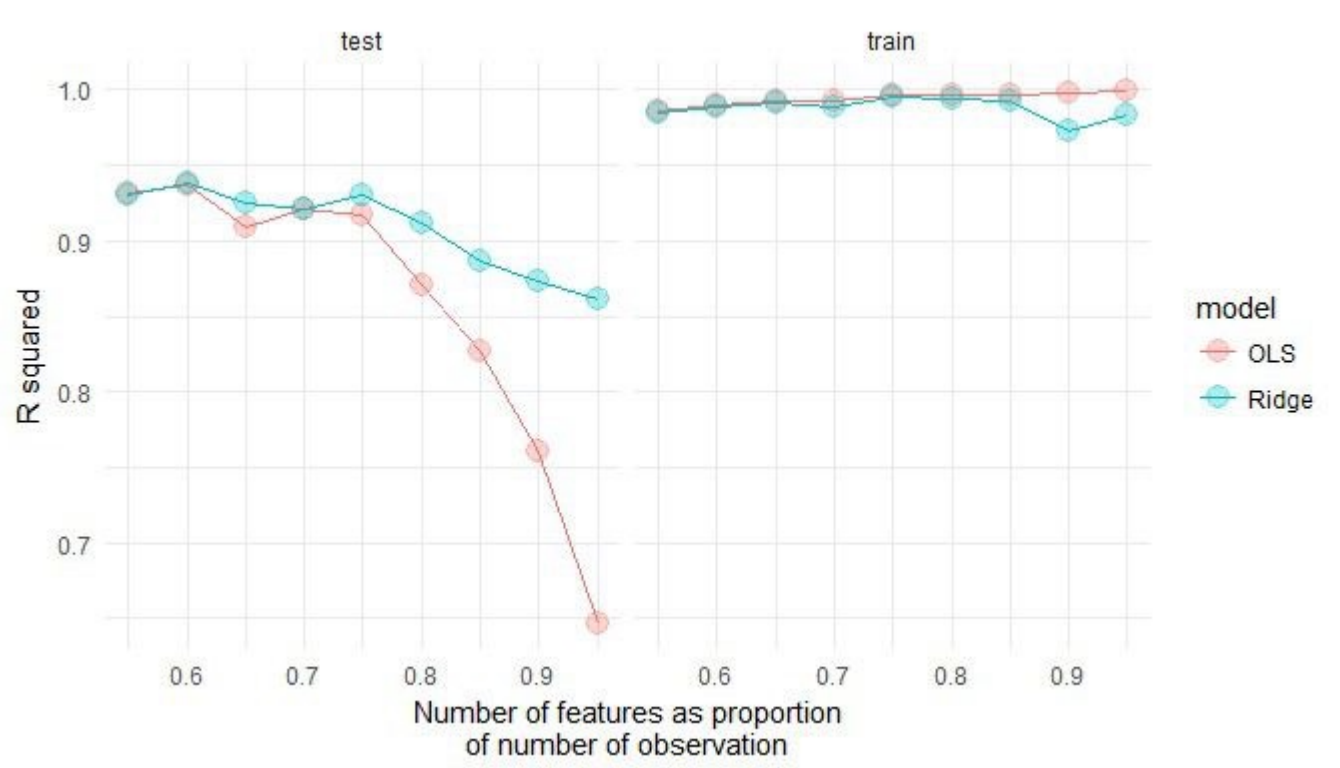


RIDGE REGRESSION (RR) VS ORDINARY LEAST SQUARES (OLS)

- Random predictors have Gaussian effects
  - $p = k*n$, where $k = 0.55,\ldots,0.95$, $n = 20,\ldots,200$
- Ridge regression does better in the test data
  - Particularly when n is small
- OLS has slightly smaller training error than RR

From:
https://drsimonj.svbtle.com/ridge-regression-with-glmnet

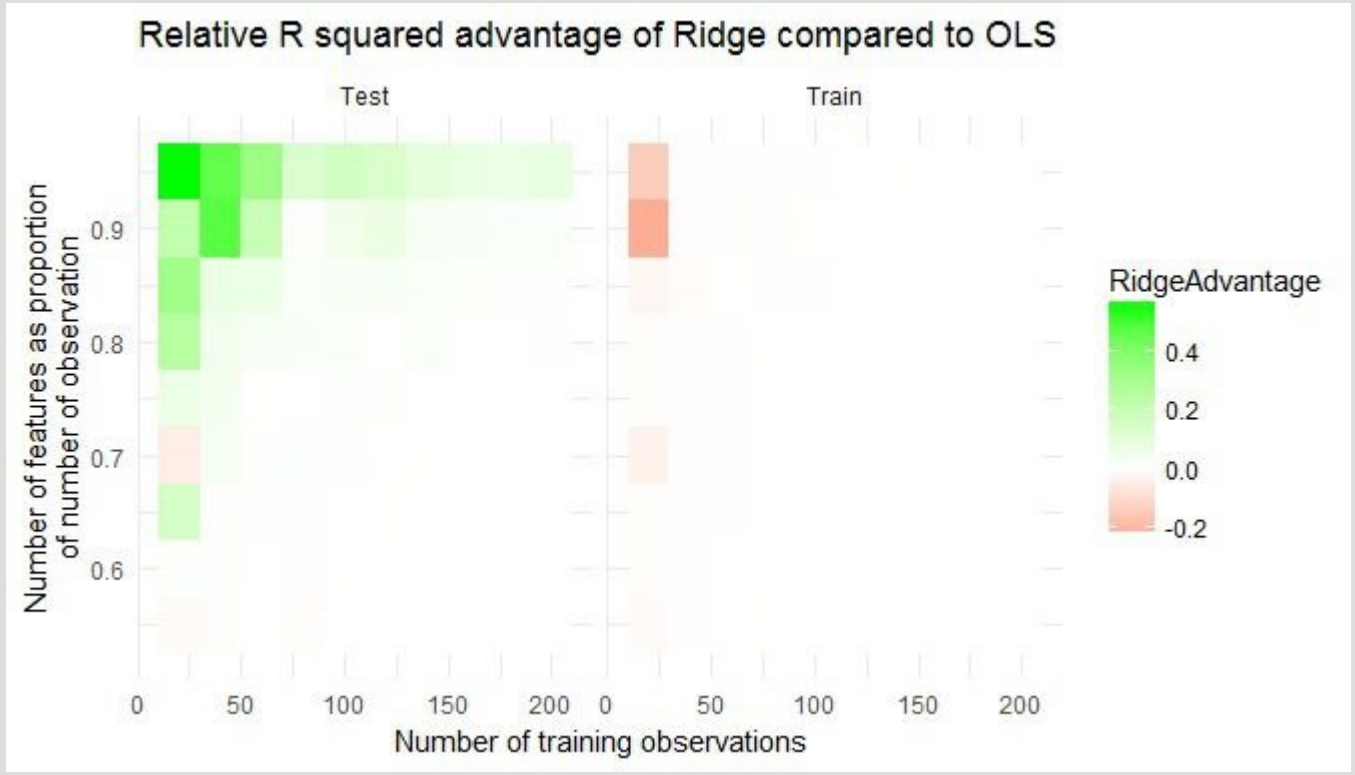
RIDGE REGRESSION VS ORDINARY LEAST SQUARES (OLS)



- Random predictors have Gaussian effects
  - p is from 55% to 95% of the sample size n
- Ridge regression does better in the test data
  - Particularly when p is large compared to n
- OLS has slightly smaller error in training data

From:
https://drsimonj.svbtle.com/ridge-regression-with-glmnet

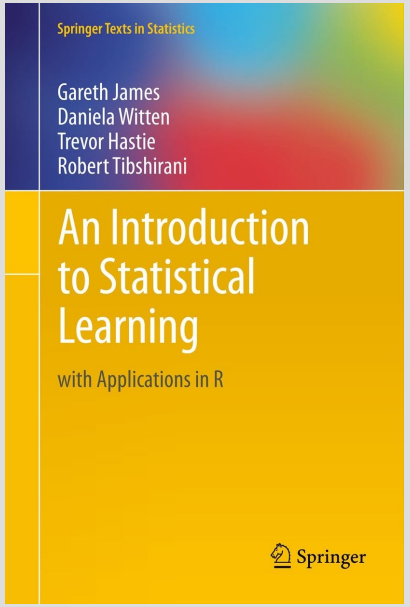
RIDGE REGRESSION VS ORDINARY LEAST SQUARES (OLS)



- Ridge regression does better in test data
  - Particularly when p is large and / or n is small
- OLS slightly better in training data
  - Overfits particularly when p is large and / or n is small

From:
https://drsimonj.svbtle.com/ridge-regression-with-glmnet

6.2 SHRINKAGE METHODS



Section 6.2

PENALIZED LIKELIHOOD FORMULATION

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2$$

Ridge regression

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

LASSO

How would you write AIC or BIC in this formulation?

CONSTRAINED MINIMIZATION FORMULATION

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s,$$

Ridge regression

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$

LASSO

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s.$$

Best subset selection. LASSO provides an efficient approximation for this

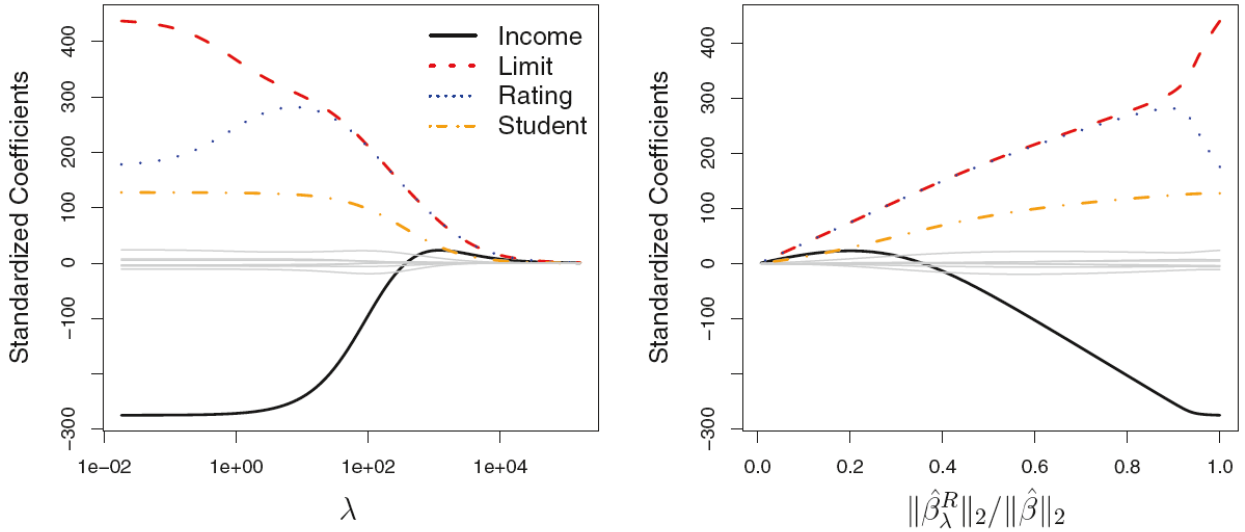


FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$.

Ridge regression:
Coefficients shrunk but never to 0
Robust estimates but not sparse models

LASSO:
Coefficient become exactly 0
at some point.
Sparse models that are
good for interpretability.

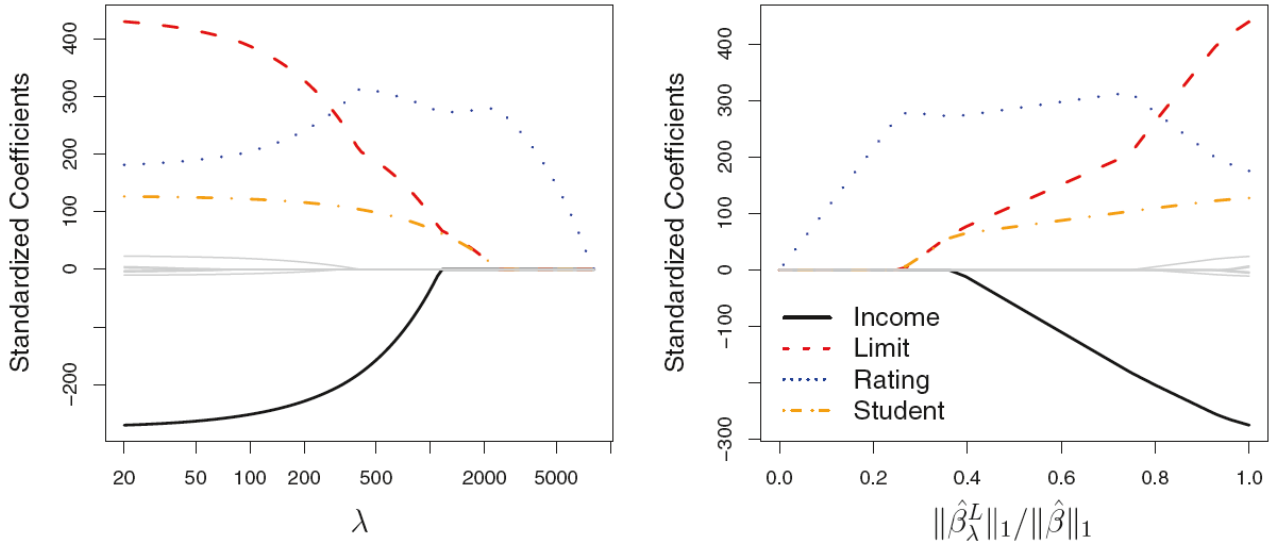


FIGURE 6.6. *The standardized lasso coefficients on the* `Credit` *data set are shown as a function of* λ *and* $\|\hat{\beta}^L_\lambda\|_1/\|\hat{\beta}\|_1$.

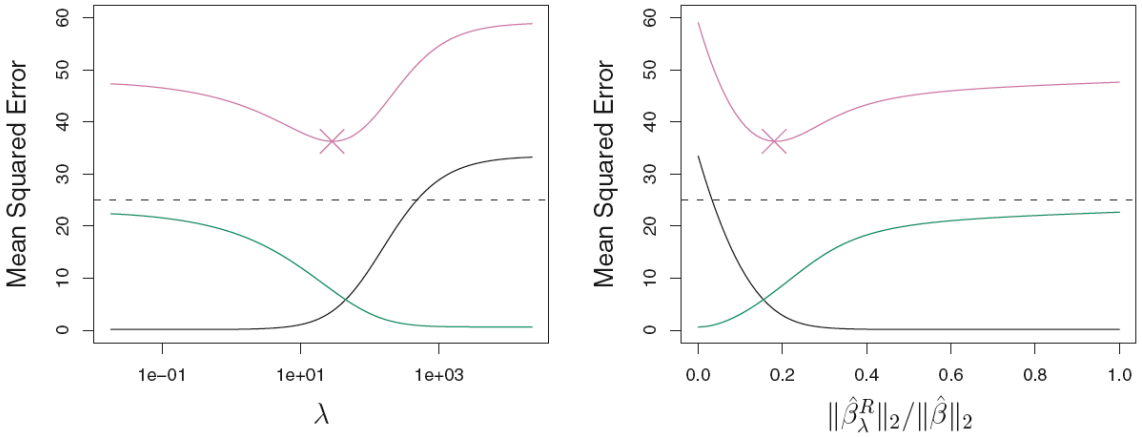


FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}^R_\lambda\|_2/\|\hat{\beta}\|_2$*. The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Data where 45 predictors all have effects

Ridge has a lower MSE than LASSO because model is not sparse and therefore LASSO is not as good as ridge regression.

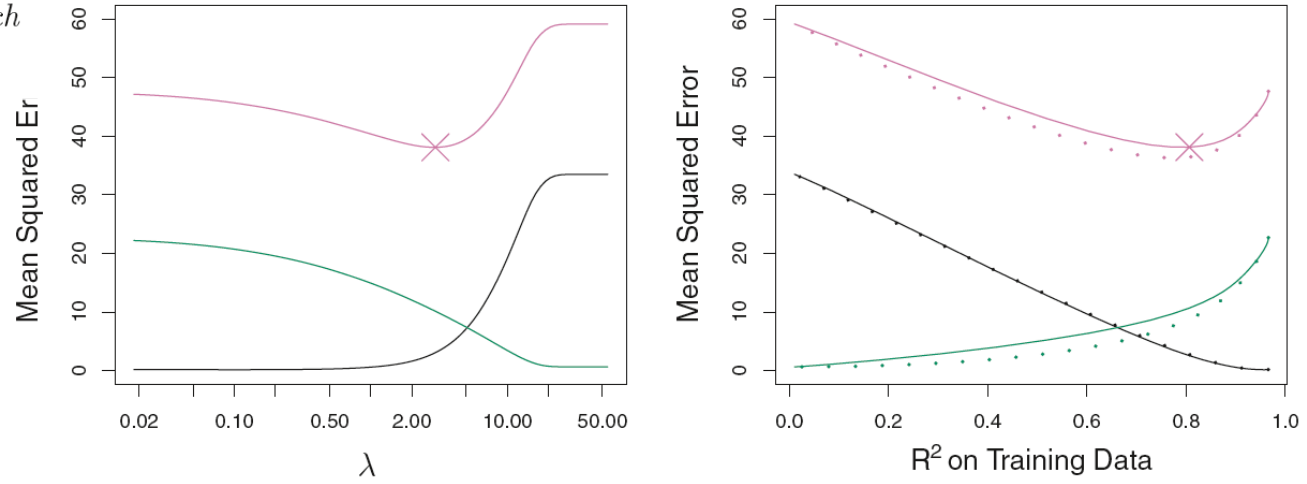


FIGURE 6.8. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso on a simulated data set.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

ADVANTAGE OF LASSO IN SPARSE MODEL

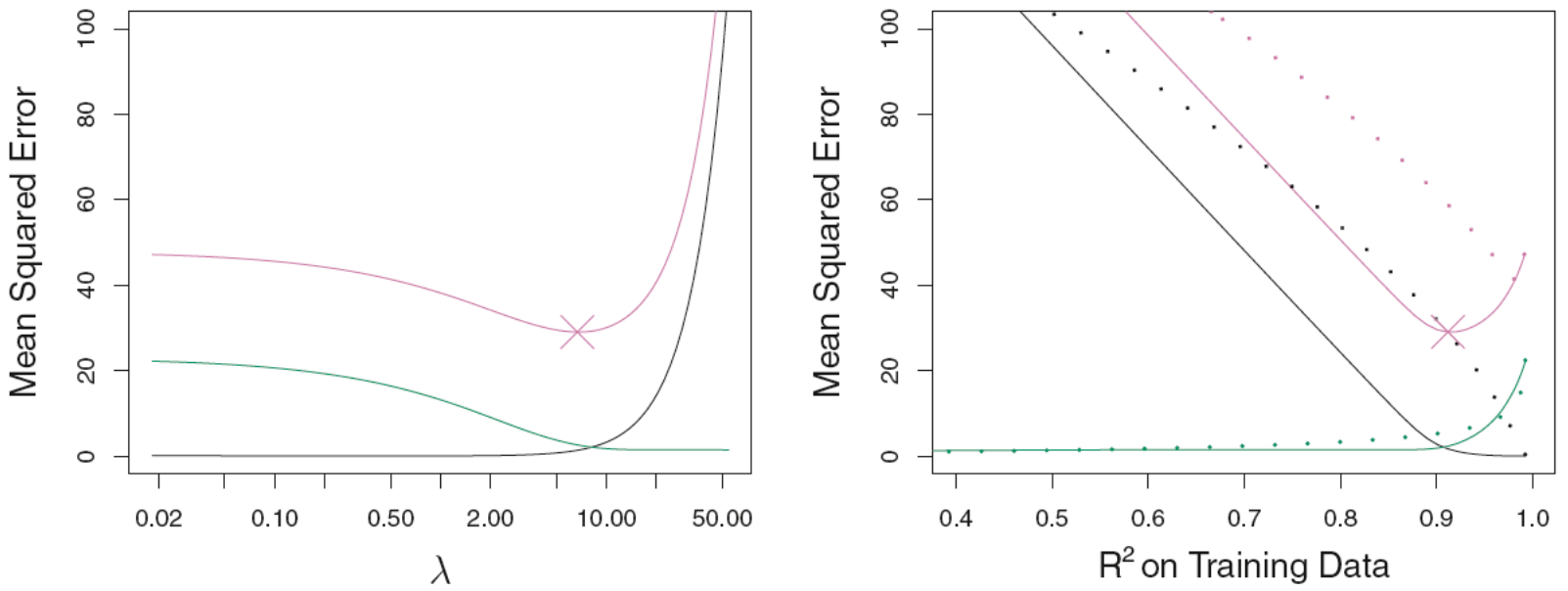


FIGURE 6.9. Left: *Plots of squared bias (black), variance (green), and test MSE (purple) for the lasso. The simulated data is similar to that in Figure 6.8, except that now only two predictors are related to the response.* Right: *Comparison of squared bias, variance and test MSE between lasso (solid) and ridge (dotted). Both are plotted against their* R^2 *on the training data, as a common form of indexing. The crosses in both plots indicate the lasso model for which the MSE is smallest.*

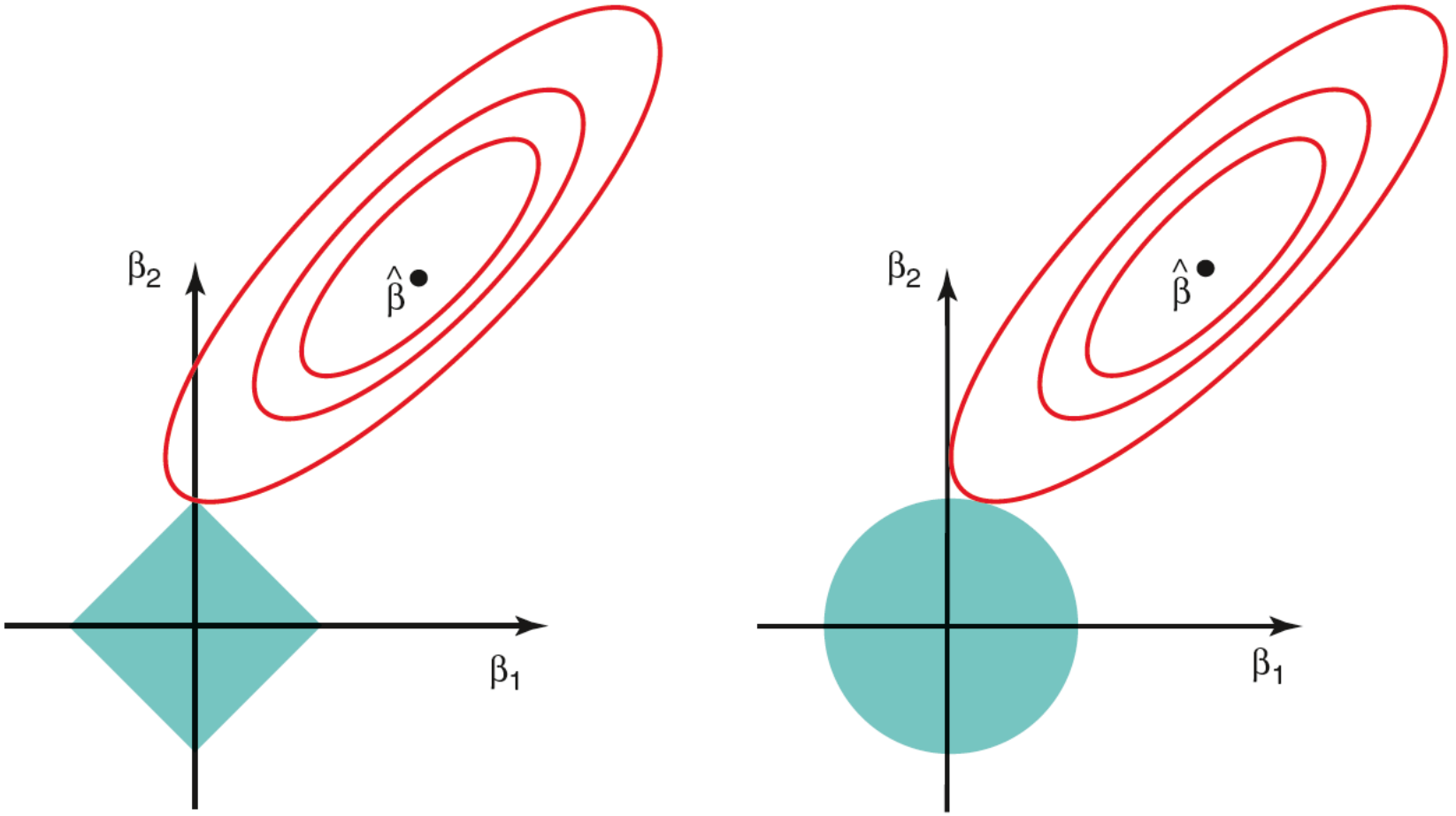
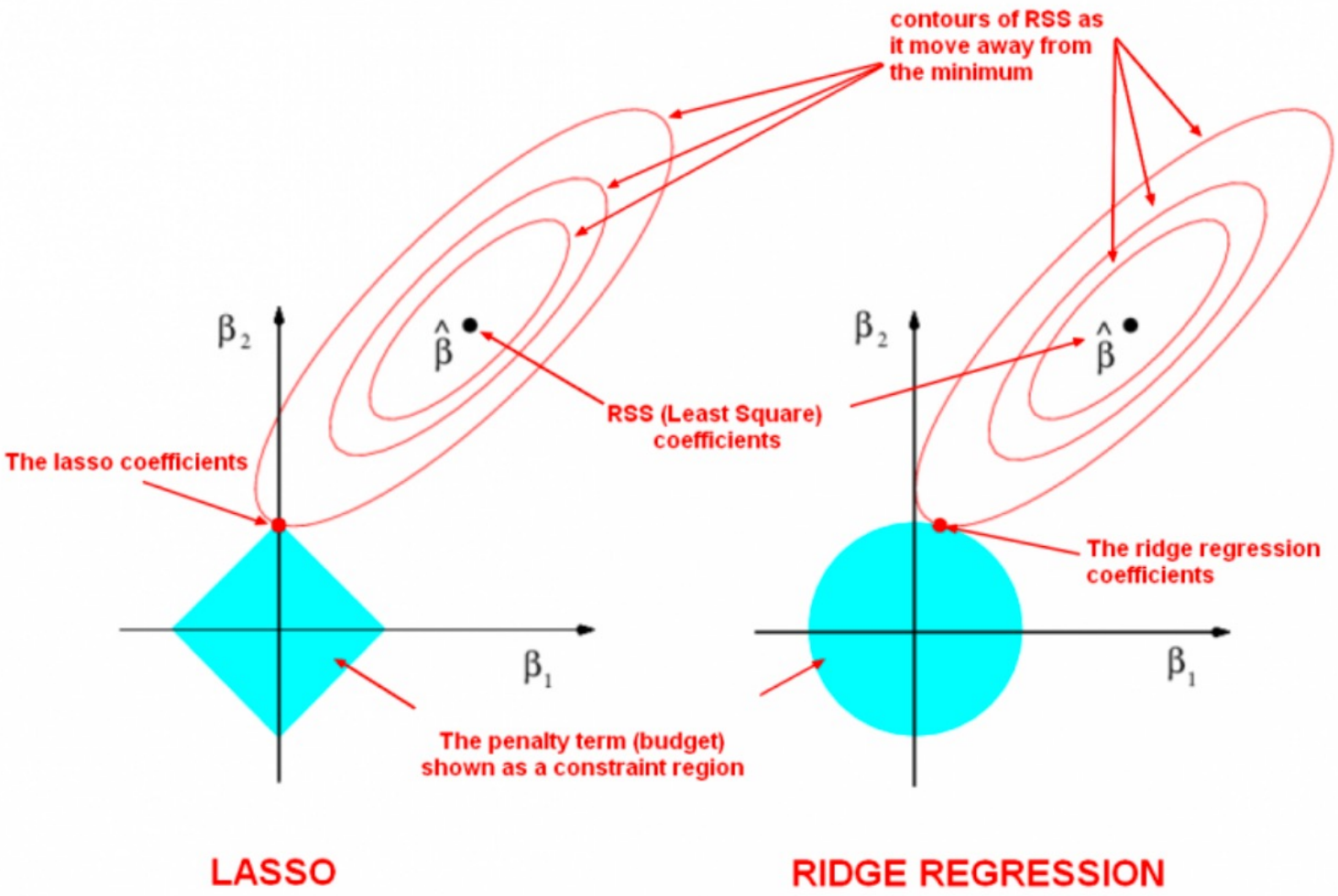


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \leq s$ *and* $\beta_1^2 + \beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Why does LASSO do variable selection and ridge does not?



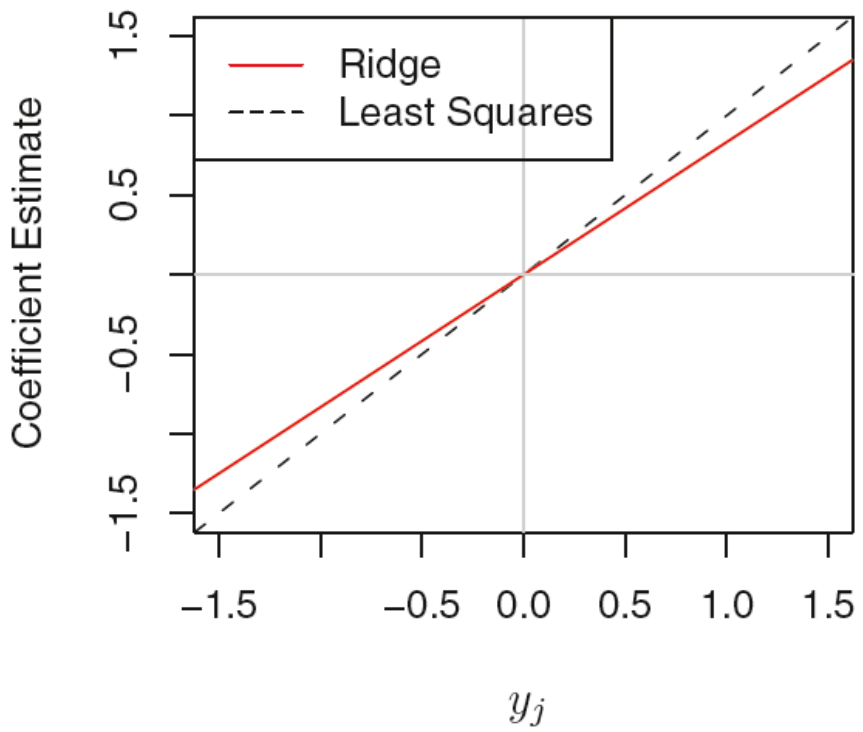
LASSO produces sparsity

In high dimensions, with LASSO, we have straight edges and corners on The coordinate axes that make a diamond. When the likelihood surface of a given value approaches the diamond, it is likely to hit the diamond at an edge or a corner where some/many coordinates are 0. This leads to some/many coefficients = 0.

RR does not produce sparsity

RR has a spherical budget region so there is no preference for the points on the coordinate axes to be the ones that hit the likelihood function at the largest value among all points in the region.

https://gerardnico.com/wiki/data_mining/lasso



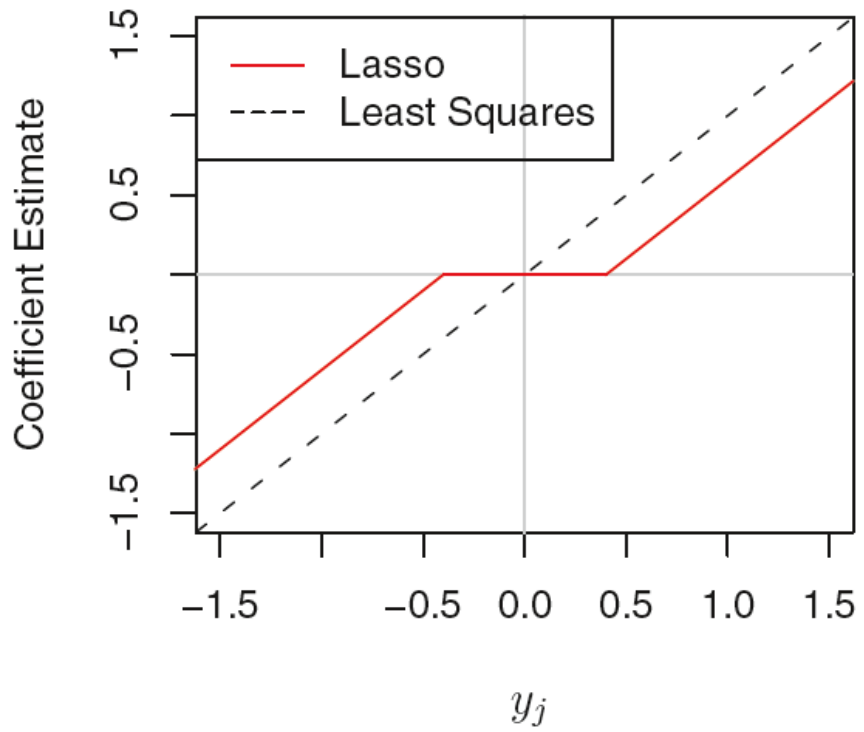


FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with $n = p$ and $\mathbf{X}$ a diagonal matrix with 1's on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

For orthogonal variables methods have simple Actions.

LASSO: Soft-thresholding, i.e. constant additive shrinkage towards 0.

Ridge: Constant multiplicative shrinkage.

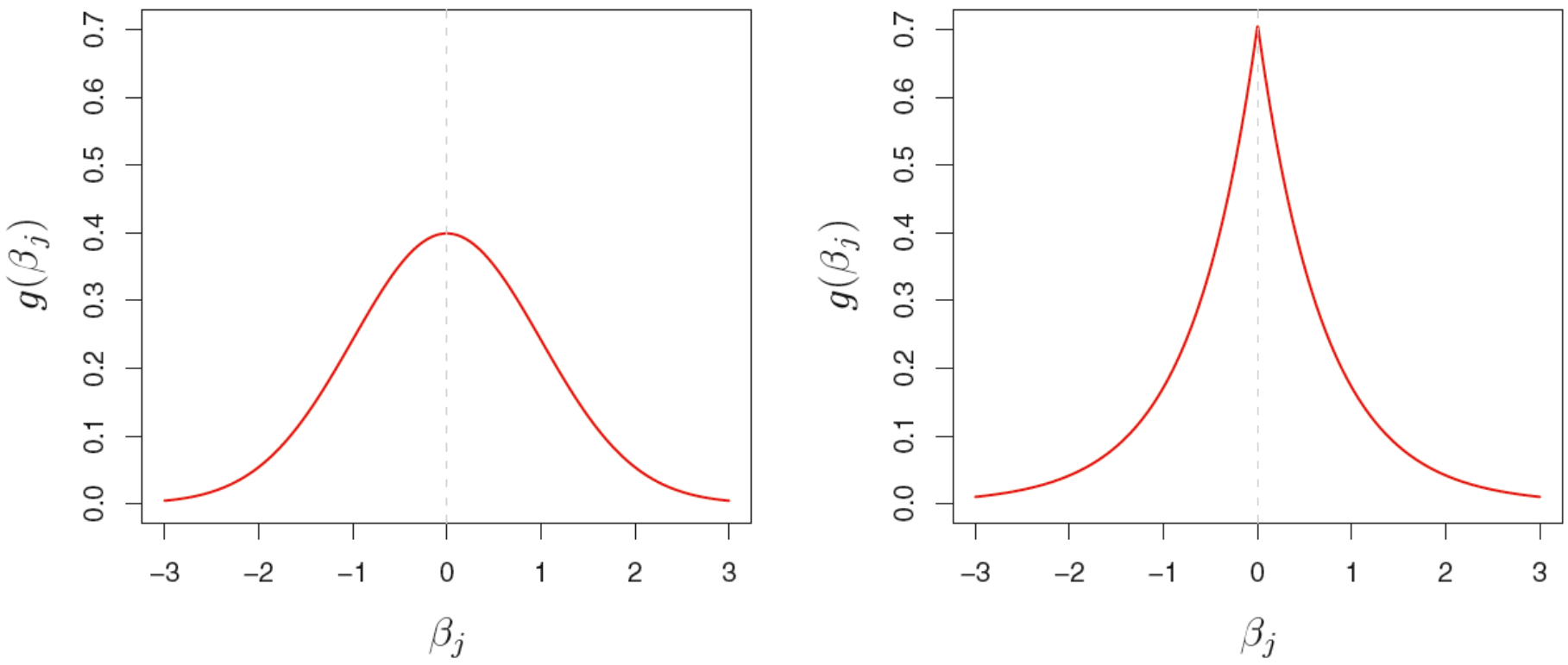


FIGURE 6.11. Left: *Ridge regression is the posterior mode for β under a Gaussian prior.* Right: *The lasso is the posterior mode for β under a double-exponential prior.*

Penalized likelihood is proportional to a posterior distribution in Bayesian statistics.

Estimates from the penalized regression are maximum a posterior values

Ridge regression uses Gaussian prior for coefficients.

LASSO uses double exponential prior for coefficients.

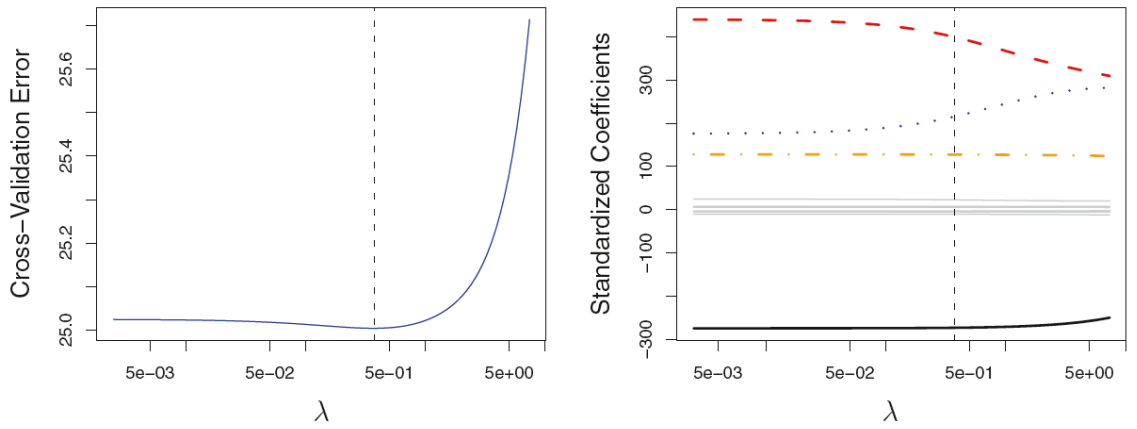


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

Cross-validation is the key to choose lambda for both methods.

Ridge regression and LASSO are flexible families of regression models that adapt their bias-variance compromise to the data through lambda value, aiming to the smallest test MSE.

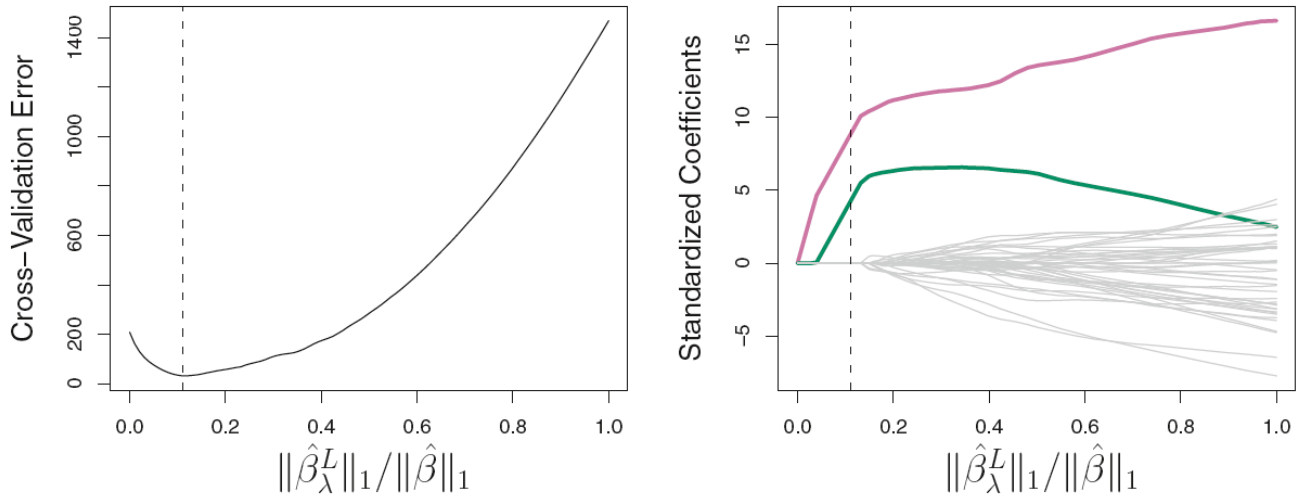


FIGURE 6.13. Left*: Ten-fold cross-validation MSE for the lasso, applied to the sparse simulated data set from Figure 6.9.* Right: *The corresponding lasso coefficient estimates are displayed. The vertical dashed lines indicate the lasso fit for which the cross-validation error is smallest.*

OTHER PENALTIES (ESL P.72-73)

- Different exponents q outside q = 1 (LASSO) and q = 2 (ridge) give different penalties
- Elastic net penalty combines LASSO and ridge penalties by linear weighting by a given parameter value α (α = 1 is LASSO, α = 0 is ridge)
- Elastic net inherits variable selection property (making some coefficients zero) from LASSO while penalties with q > 1 would not have such a property

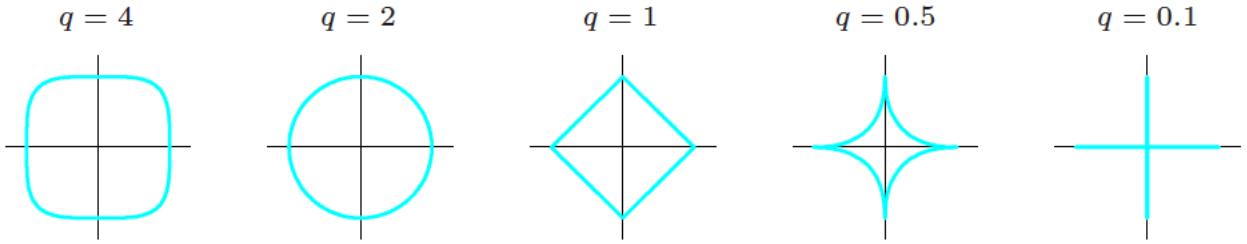
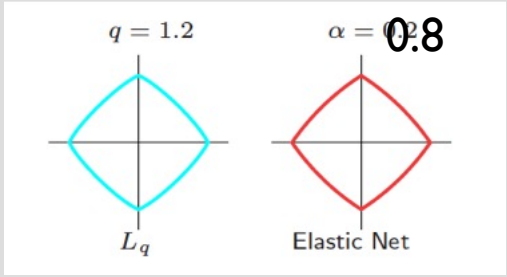


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.



$$\text{penalty}_{\text{enet}}^{\alpha}(\boldsymbol{\beta}) = \sum_{i=1}^{p}((1-\alpha)\beta_i^2 + \alpha|\beta_i|)$$

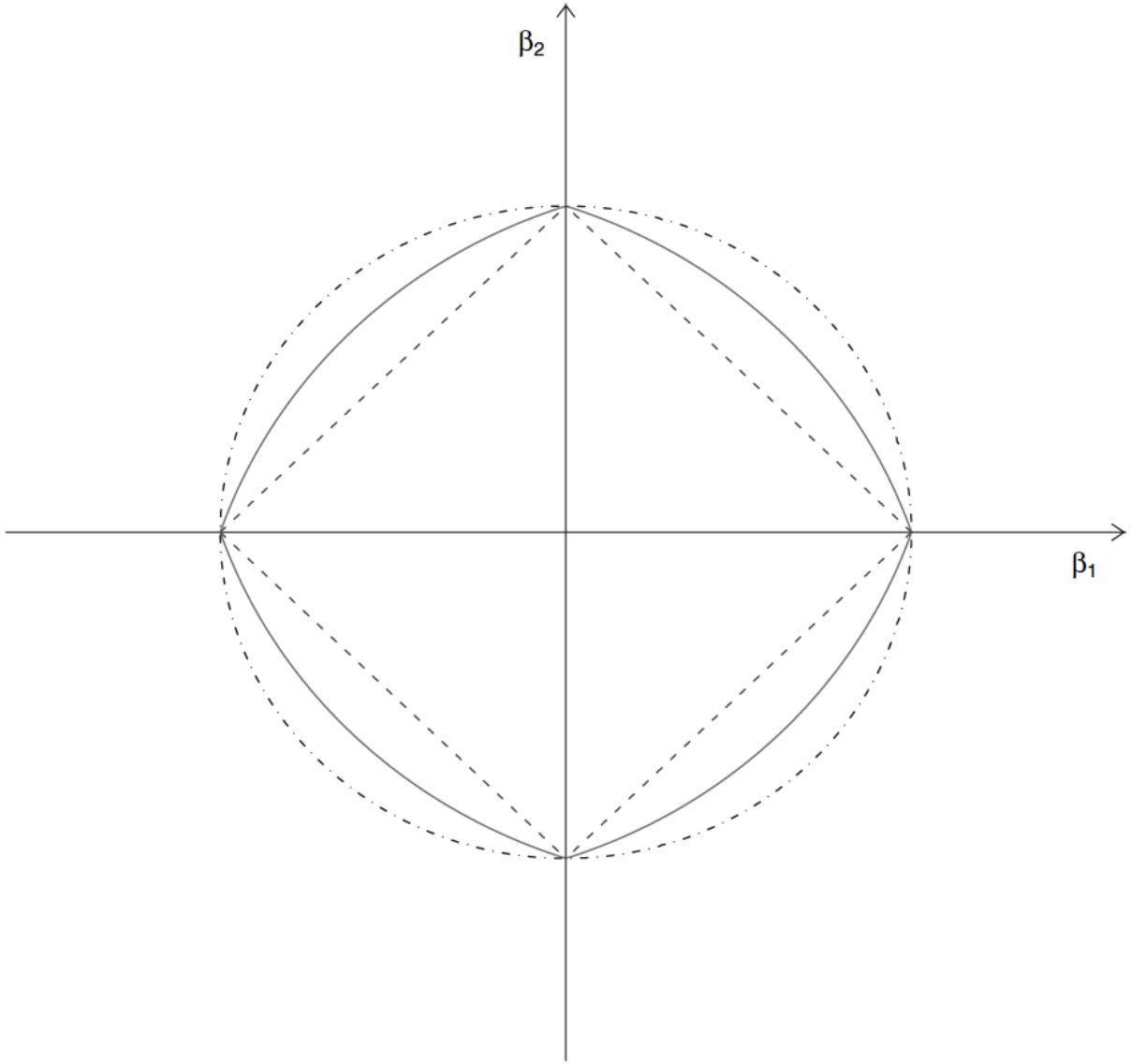


Fig. 1. Two-dimensional contour plots (level 1) (·-·-·-, shape of the ridge penalty; -------, contour of the lasso penalty; ———, contour of the elastic net penalty with $\alpha = 0.5$): we see that singularities at the vertices and the edges are strictly convex; the strength of convexity varies with α

An Introduction to `glmnet`

Trevor Hastie

Junyang Qian

Kenneth Tay

March 27, 2023

- GLMNET package
  - Does elastic **net** penalized regression for most common generalized linear models (**GLM**s)
  - Includes ridge regression (α = 0), LASSO (α =1) and linear model as special cases
  - Very fast
- Read from the beginning of the Glmnet vignette to the end of the linear regression part before you do exercises 4

https://glmnet.stanford.edu/articles/glmnet.html

Regularized Cox Regression

Kenneth Tay

Noah Simon

Jerome Friedman

Trevor Hastie

Rob Tibshirani

Balasubramanian Narasimhan

March 27, 2023

https://glmnet.stanford.edu/articles/Coxnet.html

CV FOR COX MODEL

Cox model, we compute the cross-validated deviance, which is minus twice the log partial likelihood. An issue arises in computing the deviance, since if N/K is small, there will not be sufficient observations to compute the risk sets. Here we use a trick due to van Houwelingen et al. (2006). When fold k is left out, we compute the coefficients $\widehat{\beta}^{-k}(\lambda)$, and then compute

$$\widehat{\text{Dev}}_\lambda^k := \text{Dev}[\widehat{\beta}^{-k}(\lambda)] - \text{Dev}^{-k}[\widehat{\beta}^{-k}(\lambda)]. \tag{3.29}$$

The first term on the right uses all N samples in computing the deviance, while the second term omits the fold-k samples. Finally $\text{Dev}_\lambda^{CV} = \sum_{k=1}^K \widehat{\text{Dev}}_\lambda^k$ is obtained by subtraction. The point is that each of these terms has sufficient data to compute the deviance, and in the standard cases (that is, any of the other generalized linear models), the estimate would be precisely the deviance on the left-out set.

PRE-VALIDATION

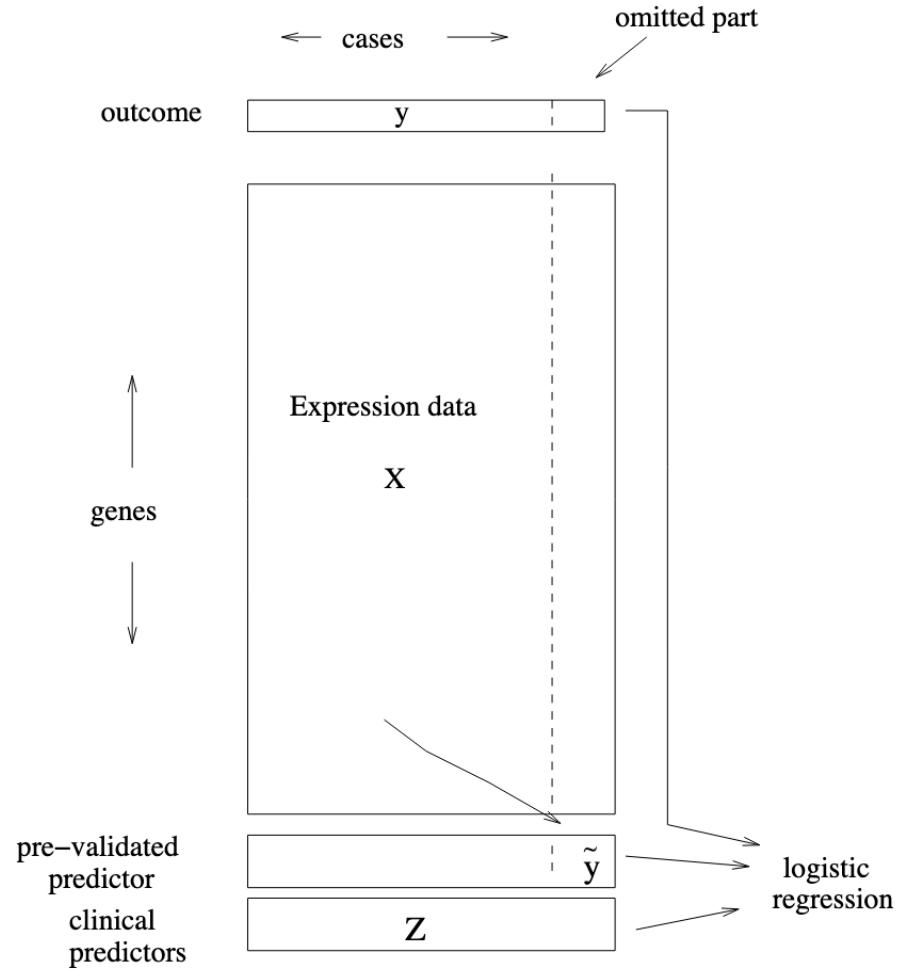


Figure 1: *A schematic of the Pre-Validation process. The cases are divided up into (say) 10 equal-sized groups. Leaving out one of the groups, a prediction rule is derived from the data of the remaining 9 groups. This prediction rule is then applied to the left out group, giving the pre-validated predictor $\tilde{y}$ for the cases in the left out group. Repeating this process for every group yields the pre-validated predictor $\tilde{y}$ for all cases. Finally, $\tilde{y}$ is included in a logistic regression model together with the clinical predictors to assess its relative strength in predicting the outcome.*

Höfling, Tibshirani:
A STUDY OF PRE-VALIDATION

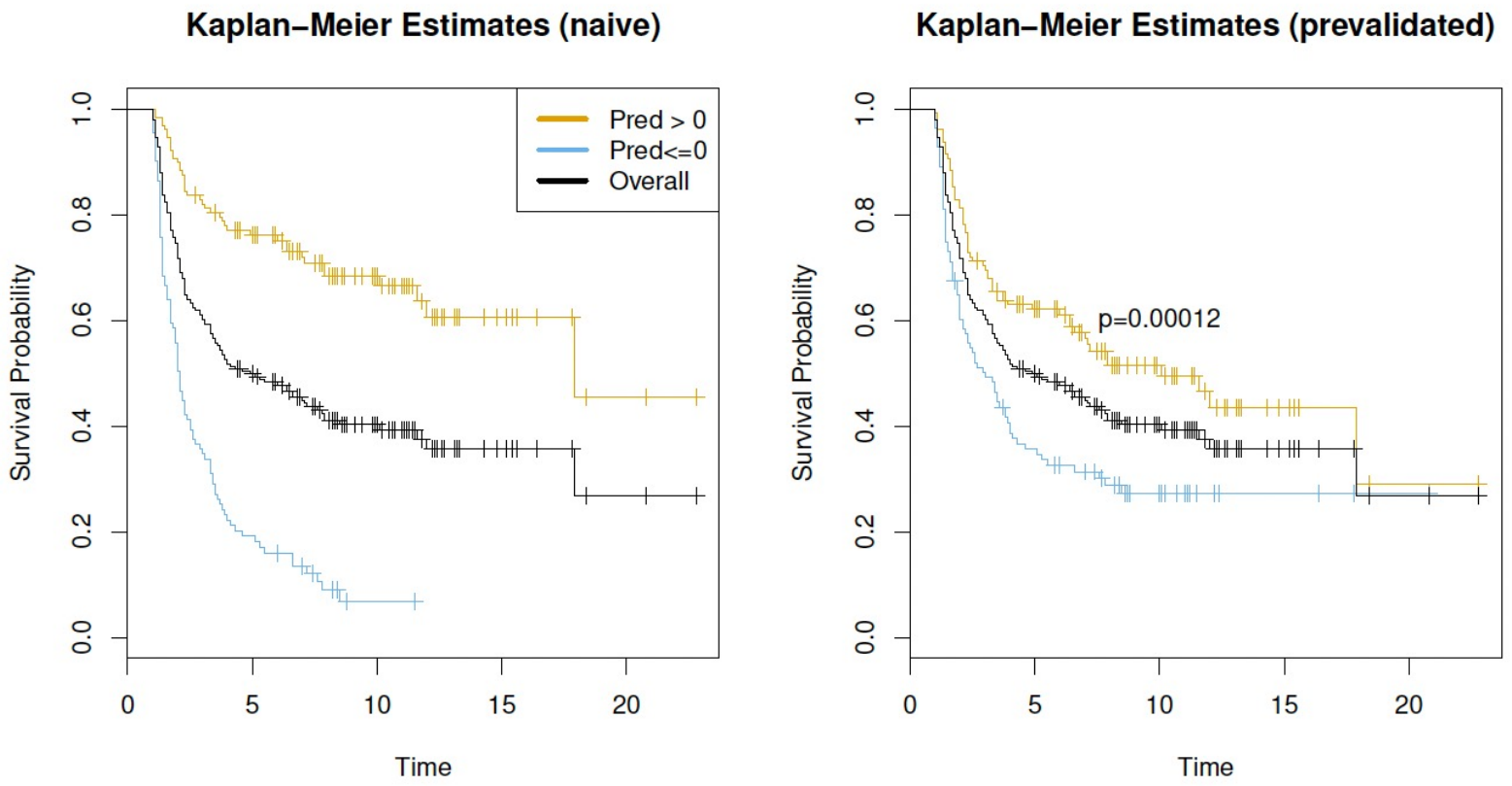
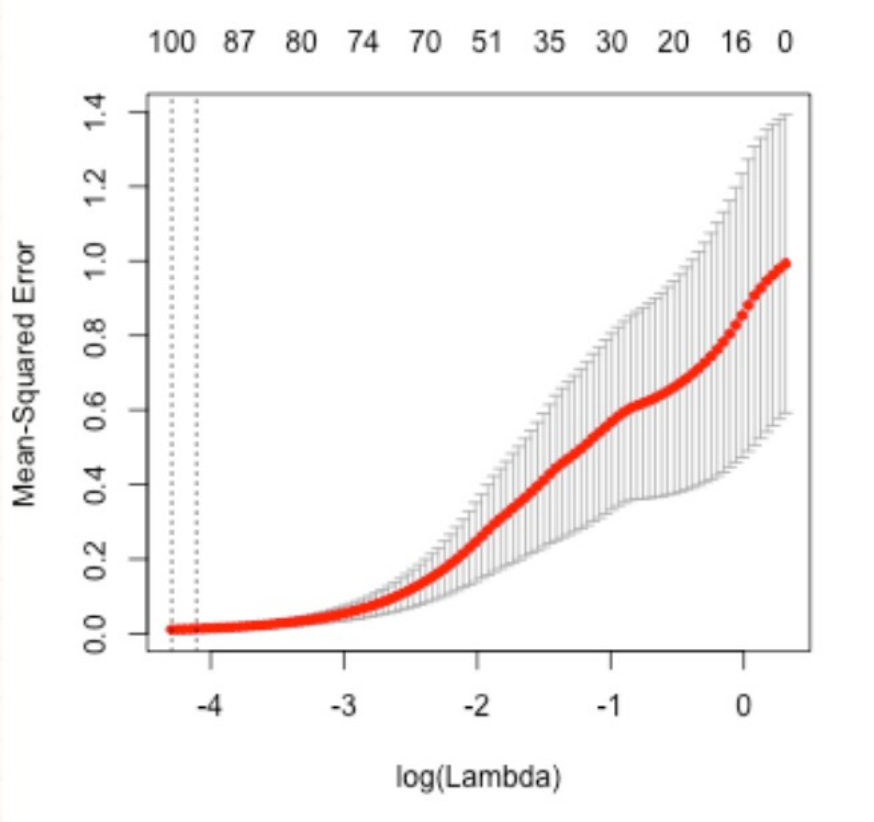
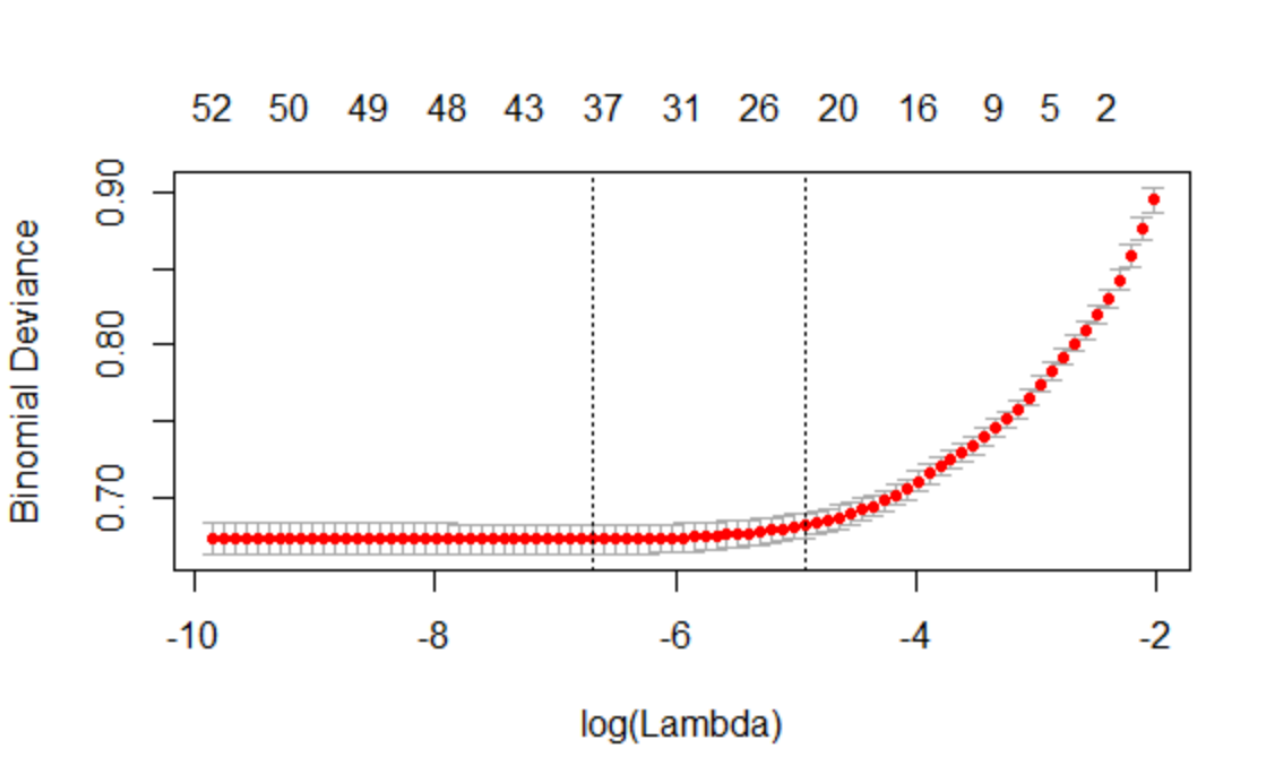


Figure 3.7 *The black curves are the Kaplan–Meier estimates of $S(t)$ for the Lymphoma data. In the left plot, we segment the data based on the predictions from the Cox proportional hazards lasso model, selected by cross-validation. Although the tuning parameter is chosen by cross-validation, the predictions are based on the full training set, and are overly optimistic. The right panel uses* prevalidation *to build a prediction on the entire dataset, with this training-set bias removed. Although the separation is not as strong, it is still significant. The spikes indicate censoring times. The p-value in the right panel comes from the log-rank test.*

- Lymphoma data
  - n = 240, p = 7399
  - Censored = 120
  - Outcome: time to death

CV.GLMNET OUTPUT





What do these plots say?

A RECENT EXAMPLE OF LASSO

Accurate Genomic Prediction Of Human Height

Louis Lello[1], Steven G. Avery[1], Laurent Tellier[1,3,5], Ana I. Vazquez[2], Gustavo de los Campos[2,4], and Stephen D.H. Hsu[1,3]

bioRxiv, Sep 18 2017

- Start with 650,000 genetic variants and 420,000 individuals with height measurements
- Use LASSO method for building the predictive model

PRE-PROCESSING

- 5 non-overlapping sets of 5k individuals each were held back from LASSO training (5 holdback sets were used for tuning lambda)
- A completely separate 5k validation set was also put aside
- A first screening based on standard univariate regression on the training set to reduce the set of candidate predictors from 645,589 to the top p = 50k and 100k by statistical significance
- Age and sex were regressed out from the outcome variable (=height) and predictors and outcome were standardized

TRAINING

- For each value of the L1 penalization λ the resulting predictor β^* is applied to the genomes of the *holdback sets* and the correlation between predicted and actual height is computed.

- It is the standard LASSO method:

and gender adjusted; both $\vec{y}$ and genotype values X are standardized).

$$\vec{\beta}^* = \underset{\vec{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} O_\lambda(\vec{y}, X; \vec{\beta}), \qquad O_\lambda(\vec{y}, X; \vec{\beta}) = \frac{1}{2}\left\|\vec{y} - X\vec{\beta}\right\|^2 + n\lambda\|\vec{\beta}\|_1, \tag{1}$$

where λ is a penalty (hyper-)parameter and the L_1 norm is defined to be the sum of the absolute values of the coefficients

$$\|\vec{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|.$$

CHOOSING LAMBDA

- A phase transition (region of rapid variation in results) occurs at roughly $10 < -log(\lambda) < 12$. The penalization is reduced until the correlation is maximized

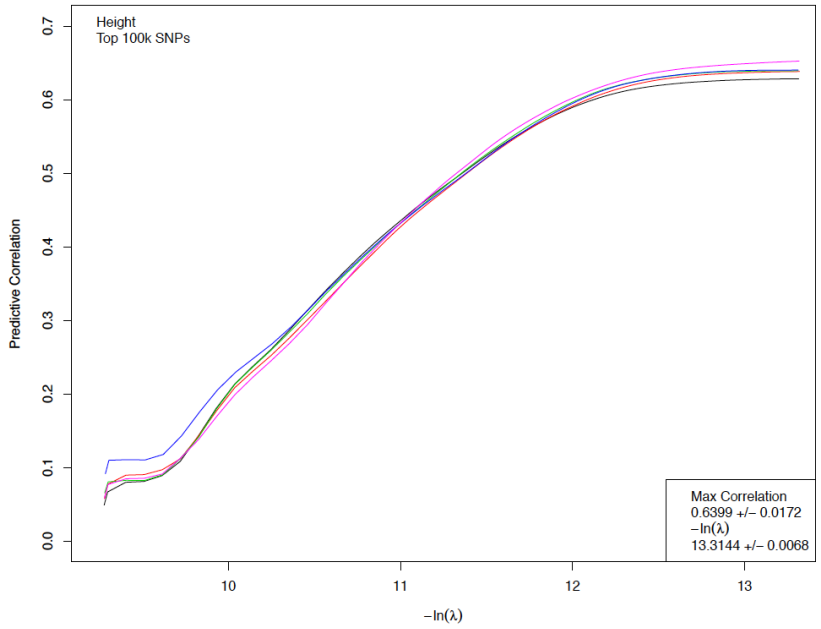


Figure 2: Correlation between actual and predicted heights as a function of L_1 penalization λ. Each line represents the training of a predictor using 453k individuals. Correlation is computed on 5k individuals not used in training.

CHOOSING LAMBDA

- The penalization is reduced until the correlation is maximized

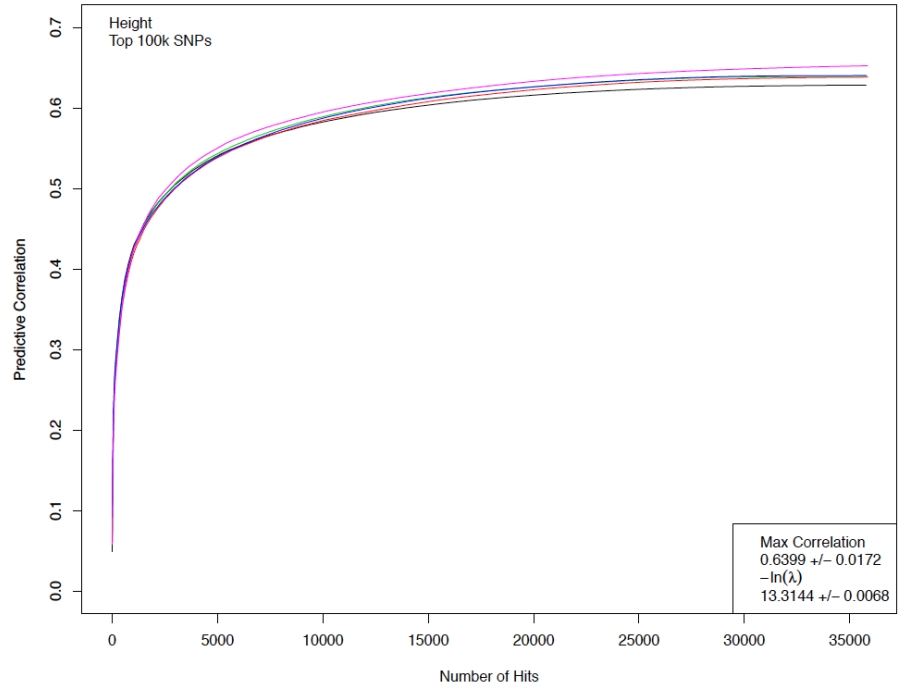
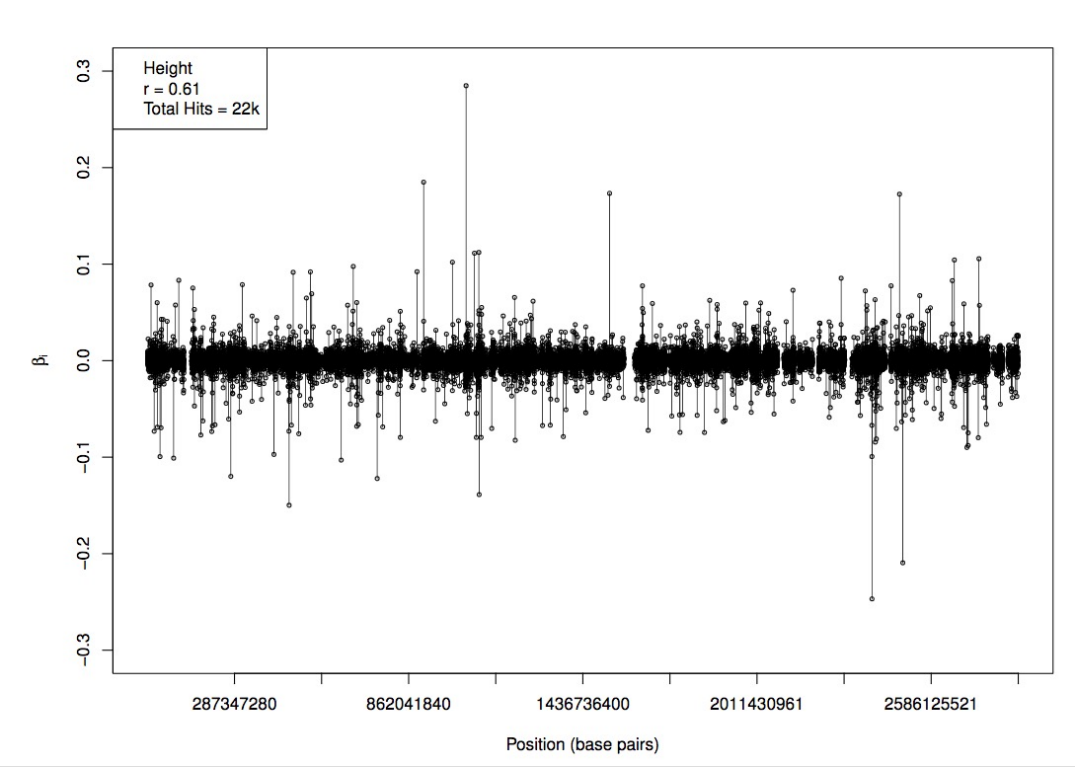


Figure 1: Correlation between actual and predicted heights as a function of the number of SNP hits activated in the predictor. While difficult to visually separate, each line represents the training of a predictor using 453k individuals. Correlation is computed on 5k individuals not used in training. The phase transition region (roughly, $10 < -\ln(\lambda) < 12$) corresponds to rapid growth in correlation on this graph, with number of hits growing from near 0 to over 5000.

IDENTIFYING RELEVANT PREDICTORS



- About 20,000 variants are identified by LASSO and each with its effect size will be used in predicting the height of a new test individual

RESULTS IN AN IN-SAMPLE VALIDATION SET (THAT WAS NOT USED IN TRAINING BUT COMES FROM THE SAME DATA SOURCE)

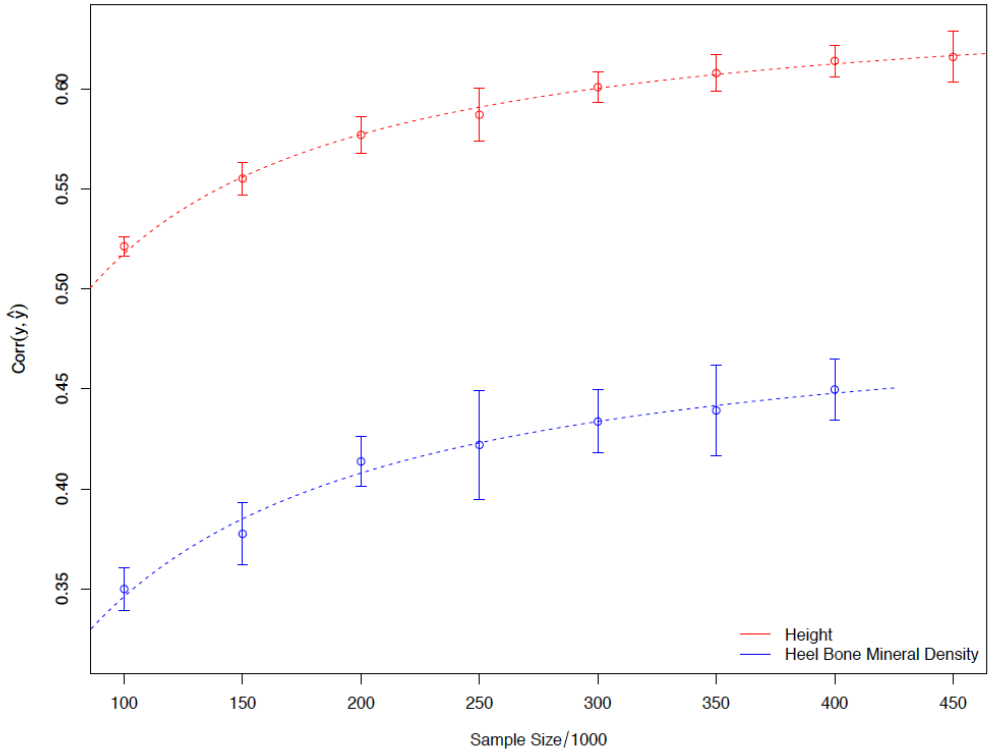


Figure 3: Correlation between predicted and actual height as number of individuals n in training set is varied. p = 50k candidate SNPs used in optimization. Fit lines of the form Corr $\sim \frac{n}{n+b}$ are included to aid visualization.

IN-SAMPLE VALIDATION

- In in-sample validation the correlation was 0.61

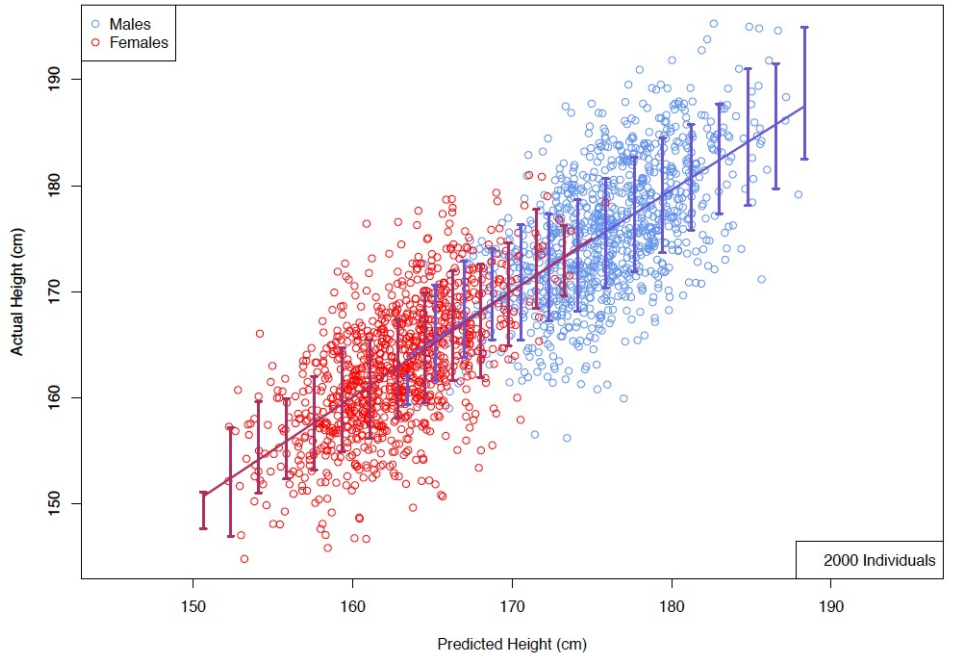


Figure 4: Actual height (cm) versus predicted height (cm) using 2000 randomly selected individuals held back from predictor optimization. Error bars indicate ±1 SD range computed using larger validation set. (Roughly equal numbers of males and females; no corrections of actual height for age or gender. See Supplement for details of predictor training.)

OUT-OF-SAMPLE VALIDATION

- In a completely separate sample the correlation dropped by ~7 percentages to ~0.54. (It already dropped from 0.61 to 0.58 in the in-sample validation set when it was restricted to the same set of predictors that were available in the out-of-sample validation set).

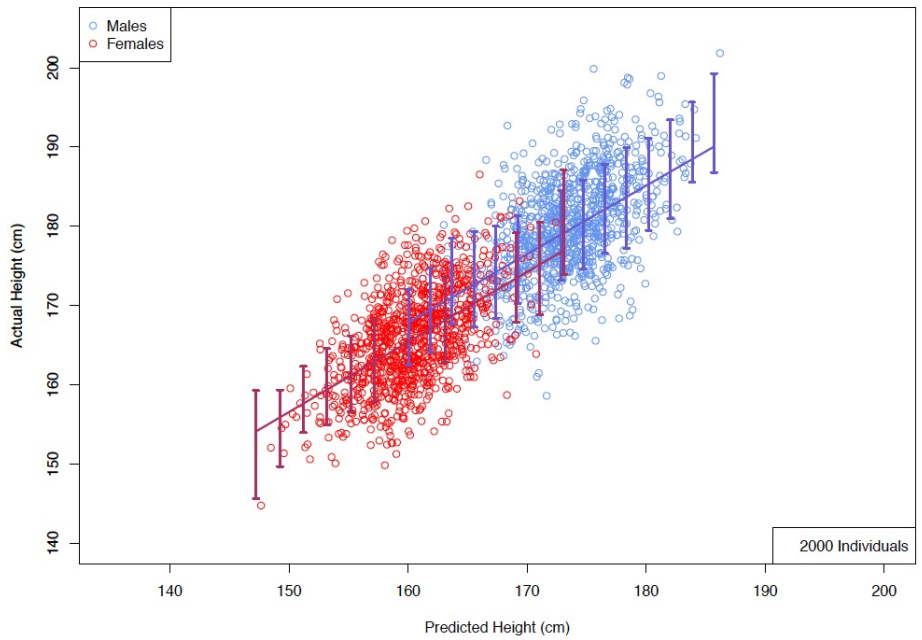


Figure 6: Actual height (cm) versus predicted height (cm) using 2000 randomly selected individuals (roughly equal numbers of M and F; no corrections for age or gender) from the ARIC dataset. Error bars indicate ±1 SD range computed using larger validation set.