“A Significance Test for the Lasso”

Lockhart R, Taylor J, Tibshirani R, and Tibshirani R

Ashley Petersen

April 25, 2013

The Tibshirani's

Motivation

- Many clinical covariates – which are important to a certain medical outcome?
- Problems with fitting model with all covariates
- Instead, choose the important variables → variable selection
- Say how important these variables are → use p-values!

Motivation

Possible variable selection techniques:

- Forward stepwise regression
- Lasso

Ways to obtain p-values:

- Forward stepwise regression: p-values from F-test used to obtain model
- Lasso: p-values from newly proposed covariance test

Being able to do proper significance testing with lasso: "bring the lasso into the mainstream" – Rob Tibshirani [1]

[1] via Andrew Gelman's blog

Forward stepwise regression

- Enter covariates into the model one at a time
- At each step choose the covariate with the largest F-statistic (smallest p-value)

$$F_k = \frac{RSS_{null} - RSS}{RSS/(n-k)}$$

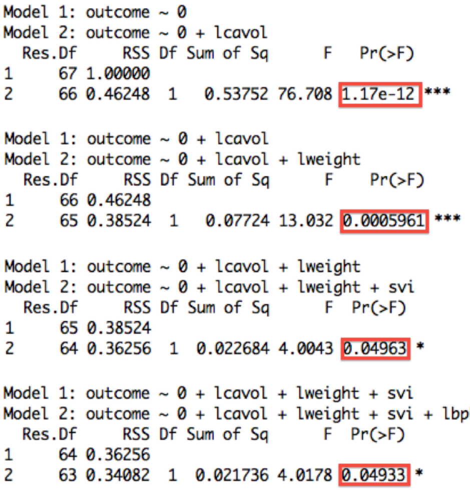
- Compare to F distribution with 1 and $n-k$ df to obtain p-value

Prostate Cancer Data

- Outcome: log PSA
- 8 covariates
- 67 observations

Example of forward stepwise regression

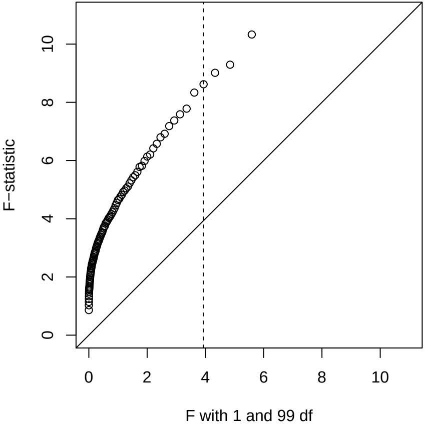
- F-test result from each step for covariate that enters the model
- Should we trust the p-values?



```
Model 1: outcome ~ 0
Model 2: outcome ~ 0 + lcavol
  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1     67 1.00000
2     66 0.46248  1   0.53752 76.708 1.17e-12 ***

Model 1: outcome ~ 0 + lcavol
Model 2: outcome ~ 0 + lcavol + lweight
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     66 0.46248
2     65 0.38524  1   0.07724 13.032 0.0005961 ***

Model 1: outcome ~ 0 + lcavol + lweight
Model 2: outcome ~ 0 + lcavol + lweight + svi
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     65 0.38524
2     64 0.36256  1  0.022684 4.0043 0.04963 *

Model 1: outcome ~ 0 + lcavol + lweight + svi
Model 2: outcome ~ 0 + lcavol + lweight + svi + lbph
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     64 0.36256
2     63 0.34082  1  0.021736 4.0178 0.04933 *
```

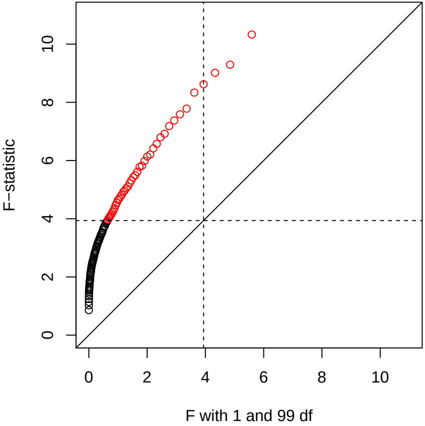
Evidence against taking those p-values seriously...

- Simulation of distribution of F-statistic for first covariate to enter model under global null ($\beta = 0$)
- $n = 100$, $p = 10$
- Type I error of 42%



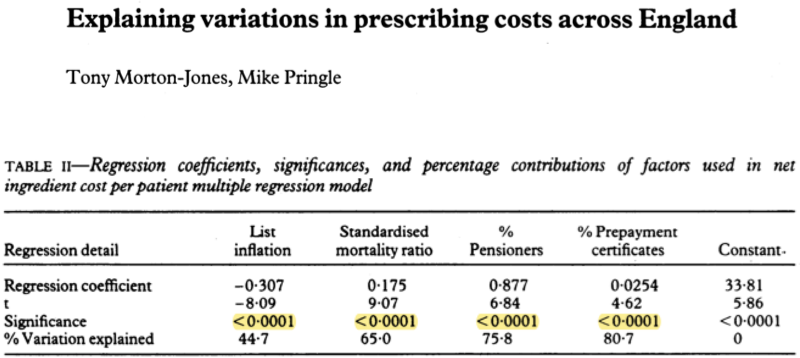
Evidence against taking those p-values seriously...

- Simulation of distribution of F-statistic for first covariate to enter model under global null ($\boldsymbol{\beta} = 0$)
- $n = 100$, $p = 10$
- Type I error of 42%



Why does this matter?

- Just look at the literature – abundance of incorrect p-values
- Much desire to do an adaptive variable selection technique and produce valid p-values



Explaining variations in prescribing costs across England

Tony Morton-Jones, Mike Pringle

TABLE II—*Regression coefficients, significances, and percentage contributions of factors used in net ingredient cost per patient multiple regression model*

Regression detail	List inflation	Standardised mortality ratio	% Pensioners	% Prepayment certificates	Constant
Regression coefficient	−0·307	0·175	0·877	0·0254	33·81
t	−8·09	9·07	6·84	4·62	5·86
Significance	<0·0001	<0·0001	<0·0001	<0·0001	<0·0001
% Variation explained	44·7	65·0	75·8	80·7	0

Lasso framework

The lasso estimator is obtained by finding $\boldsymbol{\beta}$ that minimizes

$$\frac{1}{2}\|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|,$$

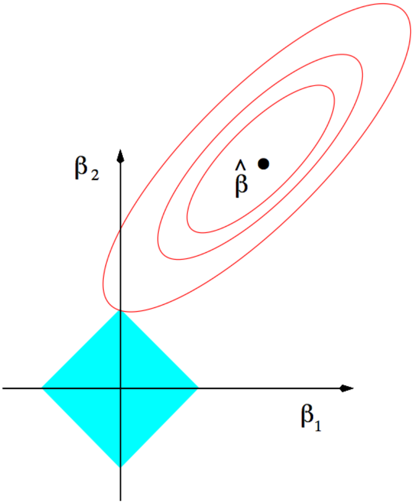
where λ is the lasso penalty. Equivalently, find $\boldsymbol{\beta}$ that minimizes

$$\frac{1}{2}\|y - X\boldsymbol{\beta}\|^2 \text{ subject to constraint } \textstyle\sum_{i=1}^{p} |\beta_i| \leq s,$$

where s is the shrinkage factor.

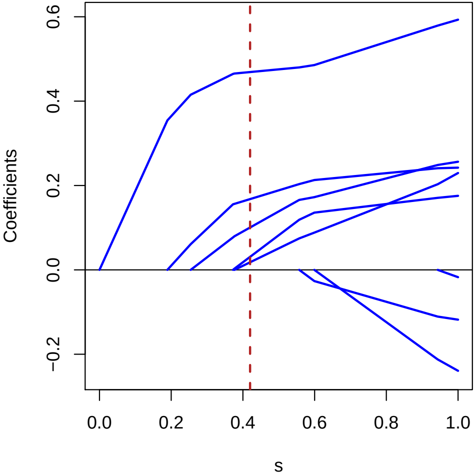
- Shrinkage and variable selection

Variable selection with lasso



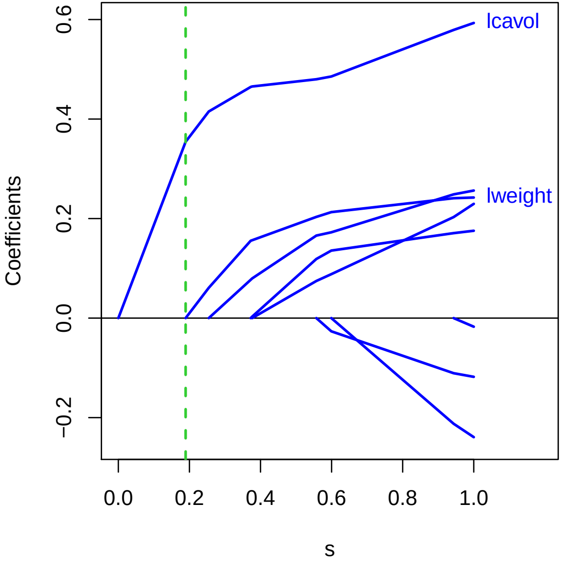
$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|y - X\boldsymbol{\beta}\|^2, \text{ subject to constraint } \sum_{i=1}^{p} |\beta_i| \leq s$$

Lasso path

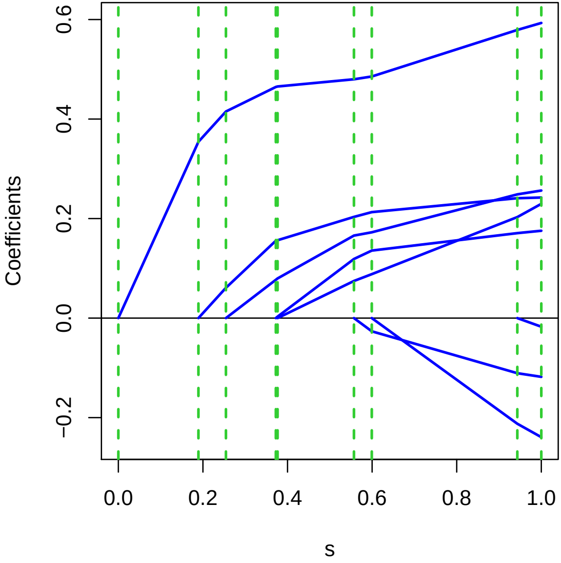


$$\hat{\boldsymbol{\beta}}_{lasso} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|y - X\boldsymbol{\beta}\|^2, \text{ subject to constraint } \textstyle\sum_{i=1}^{p} |\beta_i| \leq s$$

Obtaining p-values



Obtaining p-values



Looking back (and forward)

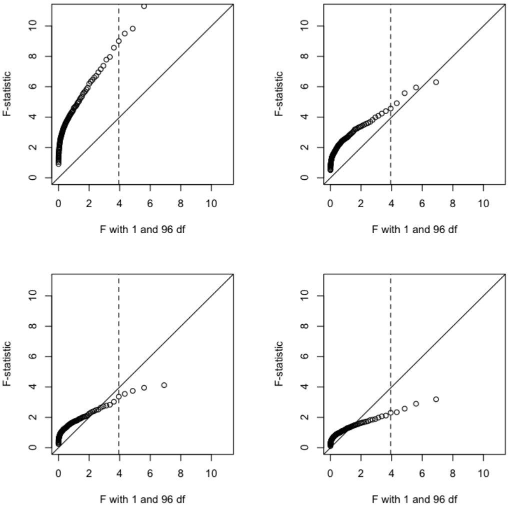
In summary:

- Working toward being able to make inferential statements in the lasso setting
- Obtain p-value for variable when it enters the lasso model
- Analogous to F-test in forward stepwise selection, but produces valid p-values

Next time:

- The test statistic and its asymptotic distribution
- Performance in finite samples using simulation

Additional simulation



Approximate type I error rates of 42%, 9%, 2%, and 0%