MA8701 Advanced methods in statistical inference and learning

W6: Statistical inference for penalized GLM methods

Mette Langaas

2/19/23

L12: 20.02.2023

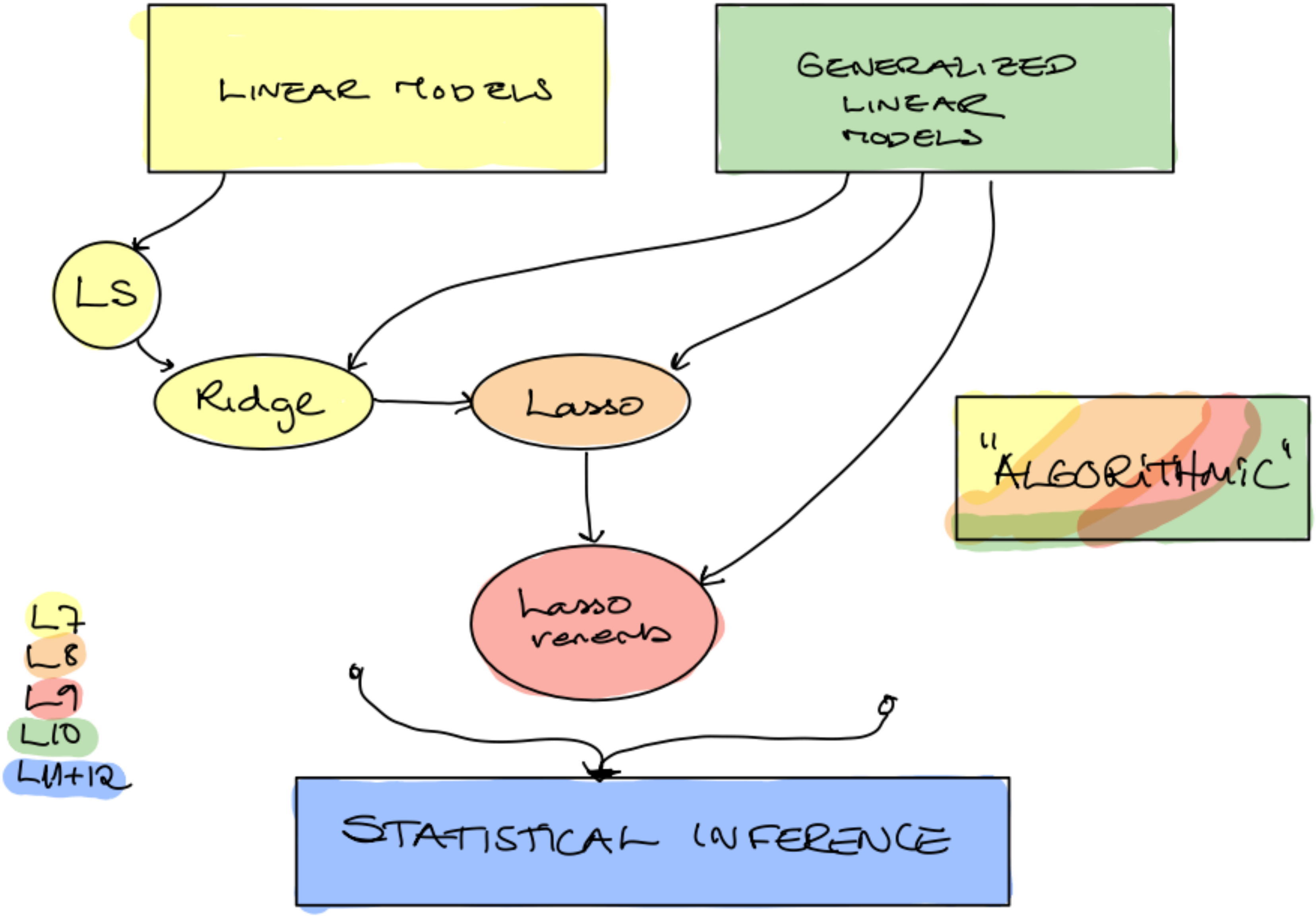


Figure 1: Overview of Part 2

Selective inference

AIM: $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

based on the results of the lasso

What is selective inference?

Selective inference is concerned with testing hypotheses suggested by the data.

data splitting

multiple testing

condition on selection

Sample splitting

What if we just split the data in two?

Lasso - linear or logistic regression

Dataset with p covariates and N observations. Divided into a training set of size aN and a test set of $(1-a)N$, where $a \in [0,1]$.

- Training data used to decide on λ using CV - gives final model where some coefficients is set to 0 and some are shrunken. (The 6 steps.)
- Test data:
  - Fit ordinary LS or GLM model with *only the non-zero lasso covariates*
  - present CI and p-values.

Group discussion: Is this ok? What is gained and what is lost?

$\div$

LOST

$a = \frac{1}{2}$

- power: use $\frac{N}{2}$ of the data

- only test H_0 for the selected covariates

  $j = 1, \dots, p$ not possible

- dependent on the split → a p-value lottery

$+$

GAINED

- valid inference

→ soon, solve this – but need knowledge from multiple testing

Single hypothesis test

$$H_0 : \beta_j = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0$$

	Not reject H_0	Reject H_0
H_0 true	Correct	Type I error
H_0 false	Type II error	Correct

Multiple hypothesis testing (m)

	Not reject H_0	Reject H_0	Total
H_0 true	u	false positive V	m_0
H_0 false	T	S	$m - m_0$
Total		R	m

For some p-value cut-off (α_{loc}) we reject R hypotheses of of m

FWER: $P(V>0) \leq \alpha$ ← "easy to work with" only involves H_0 true

prob. of no false positives

Can either find a new cut-off on the raw $\underbrace{\overset{p_j}{\text{p-values}\ (\alpha_{loc})}}$

to control FWER $\leq \alpha$

Bonferroni

$$\alpha_{loc} = \frac{\alpha}{m}$$

or $\underbrace{\text{create adjusted p-values}}$ $\tilde{p}_j$

Bonferroni $\tilde{p}_j = \min(1, m \cdot p_j)$

↑

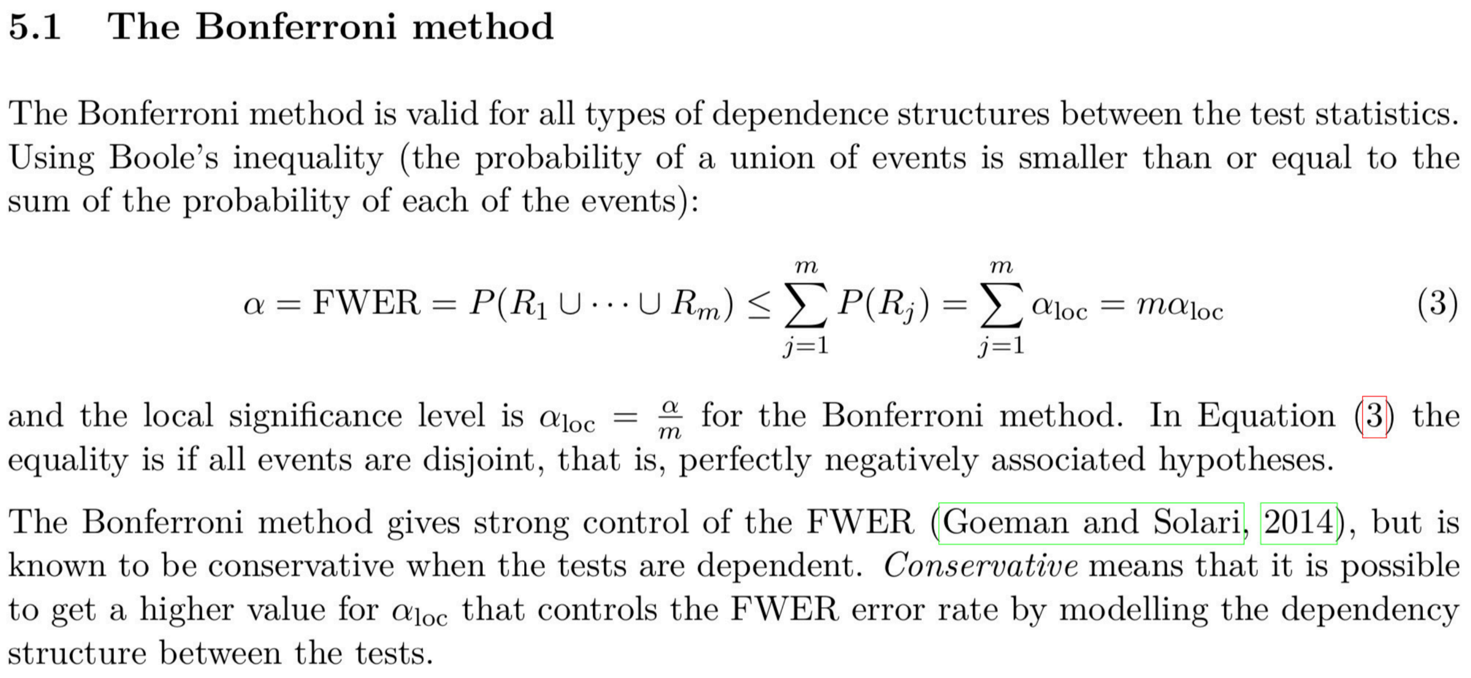
the method we will look at use this

If we reject adjusted p-values $\tilde{p}_j \leq \alpha$ then

FWER is controlled at level α for all our m hypotheses.

We have m hypothesis tests and corresponding p-values. Let us define the event R_j,

$$\begin{aligned} R_j =& \text{the } j\text{th null hypothesis is rejected} \\ =& \text{ the } p\text{-value for the } j\text{th hypothesis test is below } \alpha_{\text{loc}}. \end{aligned}$$



5.1 The Bonferroni method

The Bonferroni method is valid for all types of dependence structures between the test statistics. Using Boole's inequality (the probability of a union of events is smaller than or equal to the sum of the probability of each of the events):

$$\alpha = \text{FWER} = P(R_1 \cup \cdots \cup R_m) \leq \sum_{j=1}^{m} P(R_j) = \sum_{j=1}^{m} \alpha_{\text{loc}} = m\alpha_{\text{loc}} \tag{3}$$

and the local significance level is $\alpha_{\text{loc}} = \frac{\alpha}{m}$ for the Bonferroni method. In Equation (3) the equality is if all events are disjoint, that is, perfectly negatively associated hypotheses.

The Bonferroni method gives strong control of the FWER (Goeman and Solari, 2014), but is known to be conservative when the tests are dependent. *Conservative* means that it is possible to get a higher value for α_{loc} that controls the FWER error rate by modelling the dependency structure between the tests.

From notes
on multiple
testing

- Raw p-value, p_j, the lowest nominal level to reject the null hypothesis.
- Adjusted p-value, $\tilde{p}_j$, is the nominal level of the multiple (simultaneous) test procedure at which $H_{0j}, j = 1, \ldots, m$ is just rejected, given the values of all test statistics involved.

The adjusted p-values can be defined as

$$\tilde{p}_j = \text{inf}\{\alpha \,|\, H_{0j} \text{ is rejected at FWER level } \alpha\}$$

In a multiple testing problem where all adjusted p-value below α are rejected, the overall type I error rate (for example FWER) will be controlled at level α.

The Bonferroni method controls the FWER

Single-step methods controls for multiple testing by estimating one local significance level, α_{loc}, which is used as a cut-off to detect significance for each individual test.

The Bonferroni method is valid for all types of dependence structures between the test statistics.

The local significance level is

$$\alpha_{\text{loc}} = \frac{\alpha}{m}$$

The adjusted p-value is

$$\tilde{p}_j = \min(1, m p_j)$$

Read more here if needed: Short note on multiple hypothesis testing

High-dimensional inference

(Dezeure, Bühlmann, Meier, Meinshausen, 2.1.1 + 2.2)

- The article has focus on frequentist methods for high-dimensional inference with confidence intervals and *p*-values in linear and generalized linear models.
- We will focus on linear models.

Set-up

$Y = X\beta + \varepsilon$ $\quad E(\varepsilon)=0$ $\quad$ i.i.d $\quad$ N obs

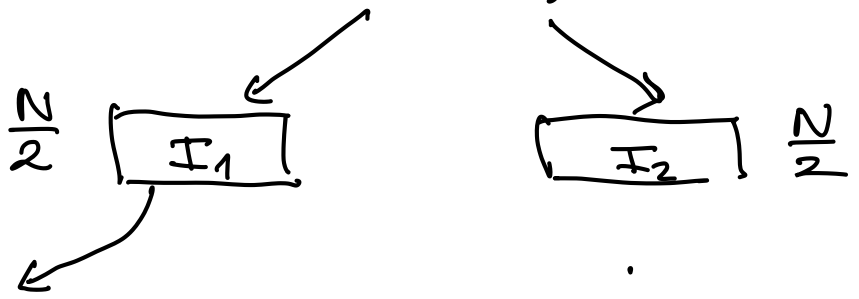
$S_0 = \{ j : \beta_j \neq 0, j = 1, \dots, p \}$ active set $\quad |S_0|$

the rest: noise variables

Aim: for each β_j, $j = 1, \dots, p$, get CI and test $H_0: \beta_j = 0$ v $H_1: \beta_j \neq 0$

SINGLE SPLIT

$I = \{1, \dots, N\}$



$I_1 \cup I_2 = I$

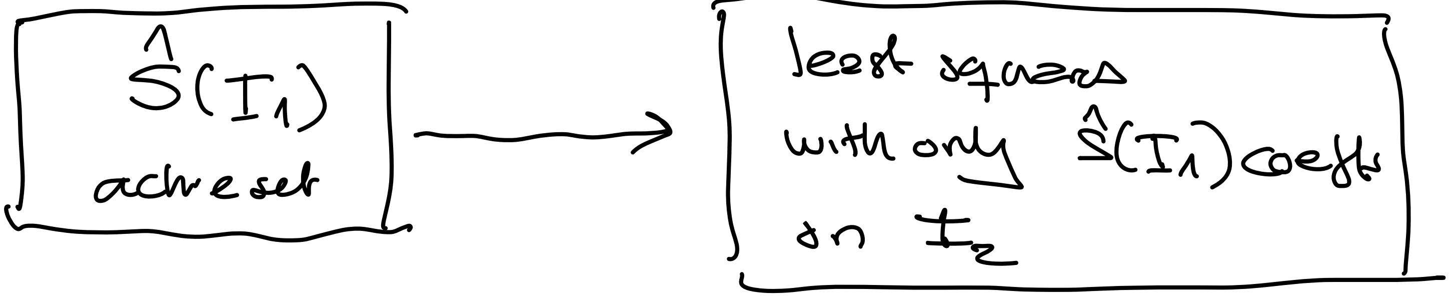
$I_1 \cap I_2 = \emptyset$

model selection with lasso

↓

$|\hat{S}(I_1)| \leq \min(N, p)$

↓



$$H_0: \beta_j = 0 \text{ vs } H_1: \beta_j \neq 0$$

$j \in \hat{S}(I_1)$: t-test to get p-value → $P_{j,raw}$

$j \notin \hat{S}(I_1)$: $P_{raw,j} = 1$

$(m \text{ hyp} = p)$

FWER adj p-values:

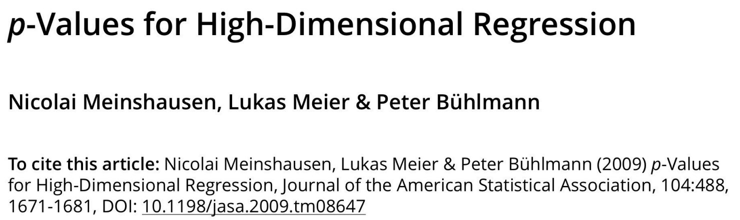
$$P_{corr,j} = \min\left(P_{raw,j}\,|\hat{S}(I_1)|, 1\right)$$

not so big correction factor?

- Controls FWER given screening property and X_{I_2} full rank

but → powerloss

Fig. 1 shown in class →



p-Values for High-Dimensional Regression

Nicolai Meinshausen, Lukas Meier & Peter Bühlmann

To cite this article: Nicolai Meinshausen, Lukas Meier & Peter Bühlmann (2009) *p*-Values for High-Dimensional Regression, Journal of the American Statistical Association, 104:488, 1671-1681, DOI: 10.1198/jasa.2009.tm08647

Multi sample splitting → repeat single sample split B times

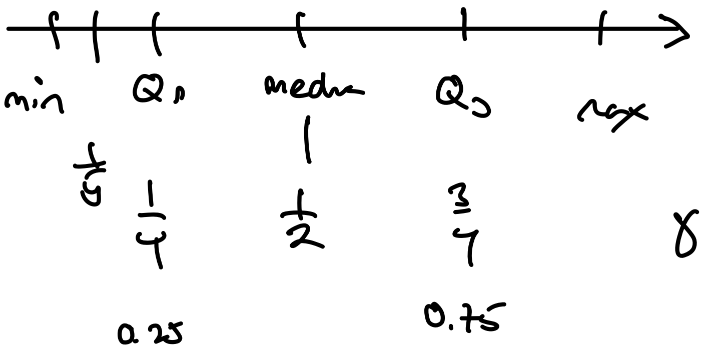
$$p_{corr,j}^{[1]}, \dots, p_{corr,j}^{[B]} \quad \text{for } j=1,\dots,p$$

How can these B adj p-values be summarized so that FWER $\leq \alpha$.

New adjusted p-value

$$Q_j(\gamma) = \min\left(1, q_\gamma\left(P_{corr,j}^{(b)}/\gamma, b=1,\dots,B\right)\right)$$

$\gamma \in (0,1)$



$\gamma = \frac{1}{2} \Rightarrow$ median $Q_j(\frac{1}{2}) = \min\left(1, \underset{b}{\text{median}}\left(2 \cdot P_{corr,j}\right)\right)$

What should γ be?

- The authors get more advanced and choose to search all γ within the interval $(\gamma_{\text{min}}, 1)$, where a common choice is $\gamma_{\text{min}} = 0.05$, to get the smallest p-value. However there is a price to pay: $(1 - \log(\gamma_{\text{min}}))$

$$P_j = \min((1 - \log(\gamma_{\text{min}}) \cdot \inf_{\gamma \in (\gamma_{\text{min}}, 1)} Q_j(\gamma)), 1)$$

for $j = 1, \ldots, p$.

Some assumptions are necessary to assure FWER control.

$\gamma_{min} = 0{,}05 \Rightarrow (1 - \log(\gamma_{min})) \approx 4 \Rightarrow$ we reject H_0 for β_j if

$$P_j \le \alpha \iff \underbrace{Q_j}_{\text{smallest over all } (\gamma_{min}, 1)} \le \frac{\alpha}{4}$$

↖ price to pay

Showed Appendix with proof of the Qj formula in class.

Confidence intervals are found by “inversion”

- from the adjusted p-values P_j
- using the duality of p-values and two-sided confidence intervals. That is, a $(1-\alpha)$ 100% CI contains values c where the p-value is below α for testing $H_0 : \beta_j = c$.
- A closed form solution involving P_j is found.
- Both single testing and multiple corrected testing CIs are found. (Appendix A.2 in article)

```
data(diabetes)
x=cbind(diabetes$x)#,diabetes$x2)
y=diabetes$y

hdires=multi.split(x=x,y=y,B=1000,fraction=0.5,
                         ci.level=0.95, model.selector=lasso.cv,
                         classical.fit=lm.pval, classical.ci=lm.c
                         return.nonaggr = FALSE, #if not adj for
                         return.selmodels=FALSE, #just to have a
                         verbose=FALSE)
dput(hdires,"hdires.dd")
```

$\frac{N}{2}$

```r
hdires=dget("hdires.dd")
names(hdires)
```

```
 [1] "pval"              "pval.corr"         "pvals.nonaggr"     "ci.level
 [5] "lci"               "uci"               "gamma.min"         "sel.mode
 [9] "method"            "call"              "clusterGroupTest"
```

```r
#summary(hdires$pvals.nonaggr) # if return.nonaggr=TRUE
hdires$gamma.min
```

```
 [1] 0.999 0.999 0.050 0.062 0.999 0.999 0.076 0.999 0.053 0.999
```

```
    adjusted pvalue    lowerCI   upperCI
age    1.000000e+00       -Inf       Inf
sex    1.000000e+00 -435.36819 106.48904
bmi    3.537003e-10  370.71236 777.71218
map    1.525473e-02   63.76631 472.25384
tc     1.000000e+00       -Inf       Inf
ldl    1.000000e+00       -Inf       Inf
hdl    5.416138e-01 -411.95903  20.84983
tch    1.000000e+00 -764.83148 204.03679
ltg    5.982750e-08  312.01305 717.79228
glu    1.000000e+00 -332.40694 242.89069
```

Summing up

What is the take home message from this "Sample splitting" story?

- relatively simple set-up $\nearrow I_1$, $\searrow I_2$ $\rightarrow P_{corr,j} \Big\} Q, P_j$

- new result to use (combine B (correlated) adj p-values using γ-quantile

- still some loss of power due to the $\frac{n}{2}$ split

- result for all p's $j = 1, \dots, p$ trouble if $|\hat{S}(I_1)| > \min(n,p)$

Inference after selection

(Taylor and Tibshirani, 2015 and HTW 6.3)

The plot

Let us leave the lasso for a while.

before 1980: small data sets, planned hypothesis to test ready before data collected, no model selection. Only fit model and look at CI and p-values.

After 1980: larger data sets and looking at data to give best model.

New challenge: *how to do inference after selection.*

This is an important topic that is not a part of ANY statistical courses at IMF.

The main question is:

- we have used a selection method (forward selection, lasso) to find potential association between covariates and response,
- with focus on interpreting the selected model: how can we assess the strength (read: CI and p-value) of these findings?

The answer includes:

- we have "cherry picked" the strongest associations, and we can thus not just report CI and p-values based on the final model - when all is done on the same data set.

In this story we now focus on *understanding how our model selection influences the inference on the final model*.

The technical solutions are of less importance, and is not presented with enough mathematical detail so that we understand the method in detail.

Remark: the single and multiple sample splitting strategy is valid.

FORWARD SELECTION (MLR)

START

$\hat\beta_1$

$\hat\beta_2$ choose $\longrightarrow$ β_{j1} $\longrightarrow$

$\vdots$

$\hat\beta_p$

single regression

with loss of gen $\beta_0=0$

$\uparrow$

$\bar y=0$

$\uparrow$ choose j_1 so that $x_{j1}^T y > x_j^T y$ for all $j \neq j_1$ } so all $x_j^T y$'s are smaller than $x_{j1}^T y$

$$\hat\beta_j = \frac{\sum_{i=1}^{N} (x_{ij} - \overset{0}{\bar x_j}) y_i}{\sum (x_{ij}-\bar x_j)^2} = \frac{1}{N-1} x_j^T y$$

let $\bar x = 0$

and $\sum x_i^2 = N-1$

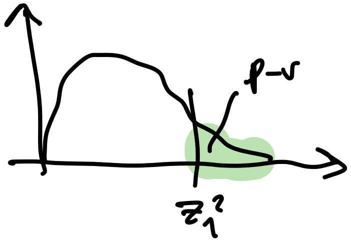
$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$

$$SD^2(\hat\beta_j) = \underset{\overset{\|}{\sum x_i^2}}{(x_j^T x_j)^{-1}}\sigma^2 = \frac{\sigma^2}{N-1}$$

if σ^2 known: $z_j = \dfrac{\frac{1}{N-1} x_j^T y}{\frac{\sigma}{\sqrt{N-1}}} = \dfrac{1}{\sigma\sqrt{N-1}} x_j^T y \sim N(0,1)$ under the H_0

$\uparrow$

$z_1^2 \sim \chi^2_1$ under H_0



β_1

$\vdots$

β_p

fit model with two β's choose $\longrightarrow$ β_{j1} β_{j2}

$$z_j^2 = \frac{(\hat\beta_j - 0)^2}{(X^TX)^{-1}_{[j,j]}\sigma^2}$$

$\swarrow$ H_0: model with β_{j1} alone better than with β_{j1} and β_{j2}

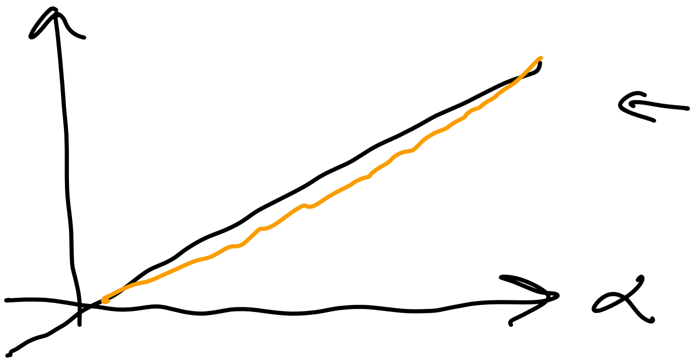
$$\frac{1}{\sigma^2}(SSE_k - SSE_{k-1}) \sim \chi^2_1$$

Alternative scenario - not selection just choose by random:

What if I add x_4 at step 1: How to check that a p-value is valid?

```
$b = 1, \dots, B$
    generate data under $H_0$ is true
    calculate p-value
```

$$\frac{\#\text{ p-values} \le \alpha}{B} \le \alpha \quad \text{for all } \alpha$$



← here the p-value has $\left(\frac{\#\text{ pvalues} \le \alpha}{B}\right) = \alpha$

$\le \alpha$

BACK TO FORWARD SELECTION

Step 1 added β_{j_1} with the largest $x_j^T y$

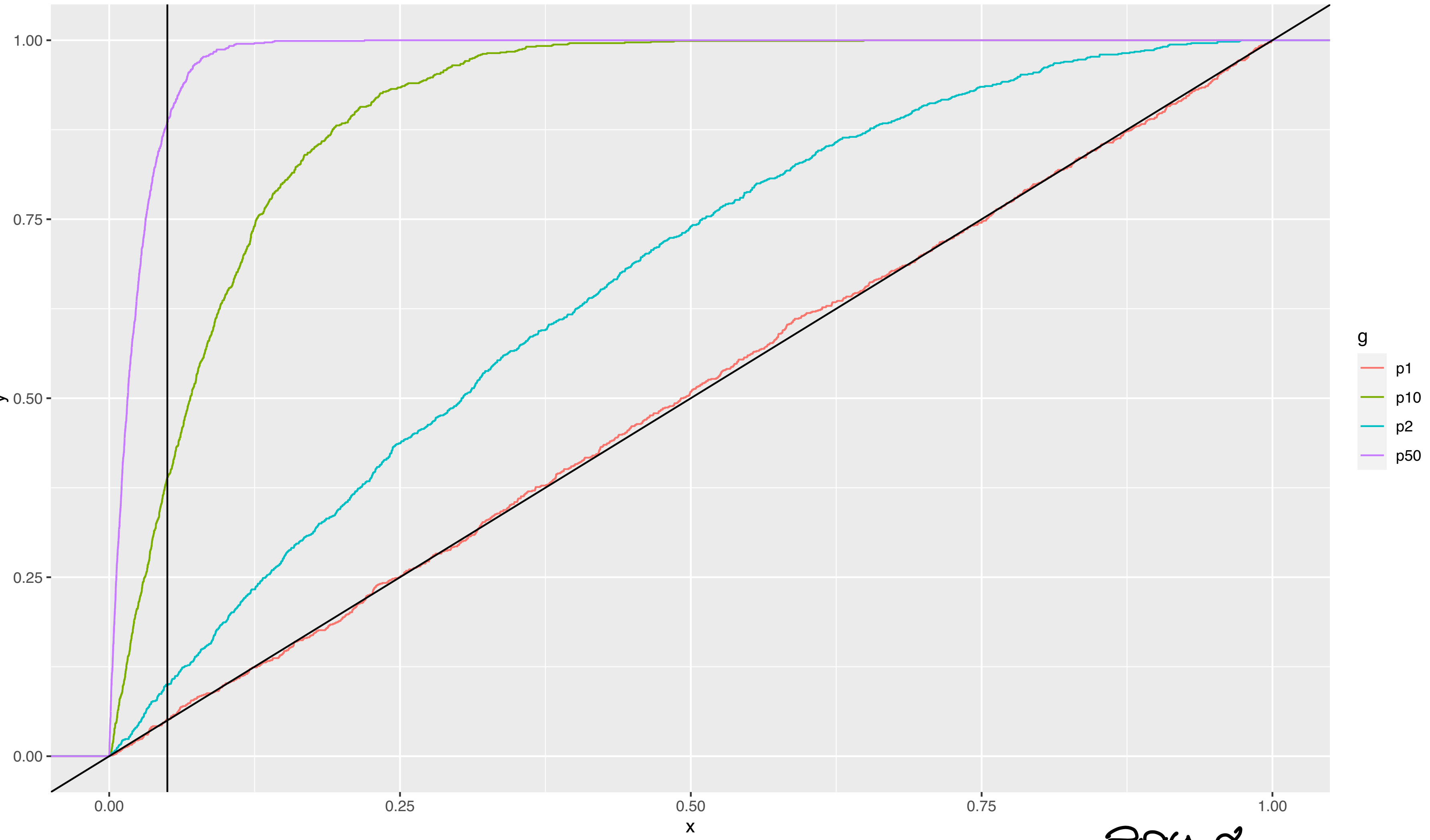
Will the distribution of $\max_j \underbrace{\frac{(x_j^T y)^2}{\sigma^2(N-1)}}_{z^2}$ be the same as the distribution of $\frac{(x_j^T y)^2}{\sigma^2(N-1)}$?

$\Rightarrow$ will the forward selection give a valid p-value?

N=100, FS in one step with $p=1$, 2, 10, 50

B=1000 sims

ECDF of p-values under the null for first step of forward selection



$\frac{\# p \leq \alpha}{B}$

actual p-value

nominal p-value

any α

$p = 1$: all good

$p > 2$: the calculated p-values are way to small

why $\underline{\max x_j^T y}$ and $\underline{x_j^T y}$ do not have the same distribution

Moving on to $k > 1$

- We would like to obtain valid ("correct") p-values for all steps, not only for $k = 1$.
- Monte Carlo solution would be elaborate.

The method used in the article is to calculate a p-value for the covariate at step k by conditioning on the fact that the strongest $k - 1$ predictors in this sequential set-up has already been chosen.

The p-value to be calculated at step k would be dependent on the number of covariates p.

We now change focus and look at the distribution of the estimated regression coefficient for the covariate added at step k, because that can be used to construct both a CI for the coefficient and a p-value for testing if the coefficient is different from zero.

The polyhedral result

(for details consult HTW 6.3 or articles references to in the Taylor and Tibshirani article)

Distribution for regression coefficient:

- Assume that we are at some step k, and that $k-1$ covariates are in the model.
- We have found the new covariate to include, and fitted the model with the k covariates.
- Standard theory tells us that the estimator $\hat{\beta}$ for covariate k is unbiased and follows a normal distribution with some variance τ^2.

$$\hat{\beta} \sim N(\beta, \tau^2)$$

But, this is given that we only had these k covariates available at the start. We will instead *condition on* selection event.

It turns out that the selection event can be written in a *polyhedral form* $Ay \leq b$ for some matrix A and some vector b.

At each step of the forward selection we have a competition among all p variables, and the A and b is used to construct the competition.

In the following case the region is nicely specified

orthogonal regression $X^TX = I$

$y_i \sim N(0,1)$, X independent of y

$x_{j_1}^T y$ largest possible → we need to only look at

the set $\{y : x_{j_1}^T y \geq \pm x_j^T y$ for all $j \neq j_1\}$

Then we find the distribution of $x_{j_1}^T y$ conditioned on

$$\underset{\substack{\uparrow \\ V^{lo}}}{\max_{j \neq j_1} |x_j^T y|} \le x_{j_1}^T y \le \underset{\substack{\uparrow \\ V^{up}}}{\infty} \qquad Ay \le b$$

The correct distribution of the estimator $\hat{\beta}$ for covariate now has a *truncated normal distribution*

$$\hat{\beta} \sim TN^{c,d}(\beta, \tau^2)$$

i.e. the *same* normal distribution, but scaled to lie within the interval (c, d).

The limits (c, d) depends on both the data and the selection events that lead to the current model.

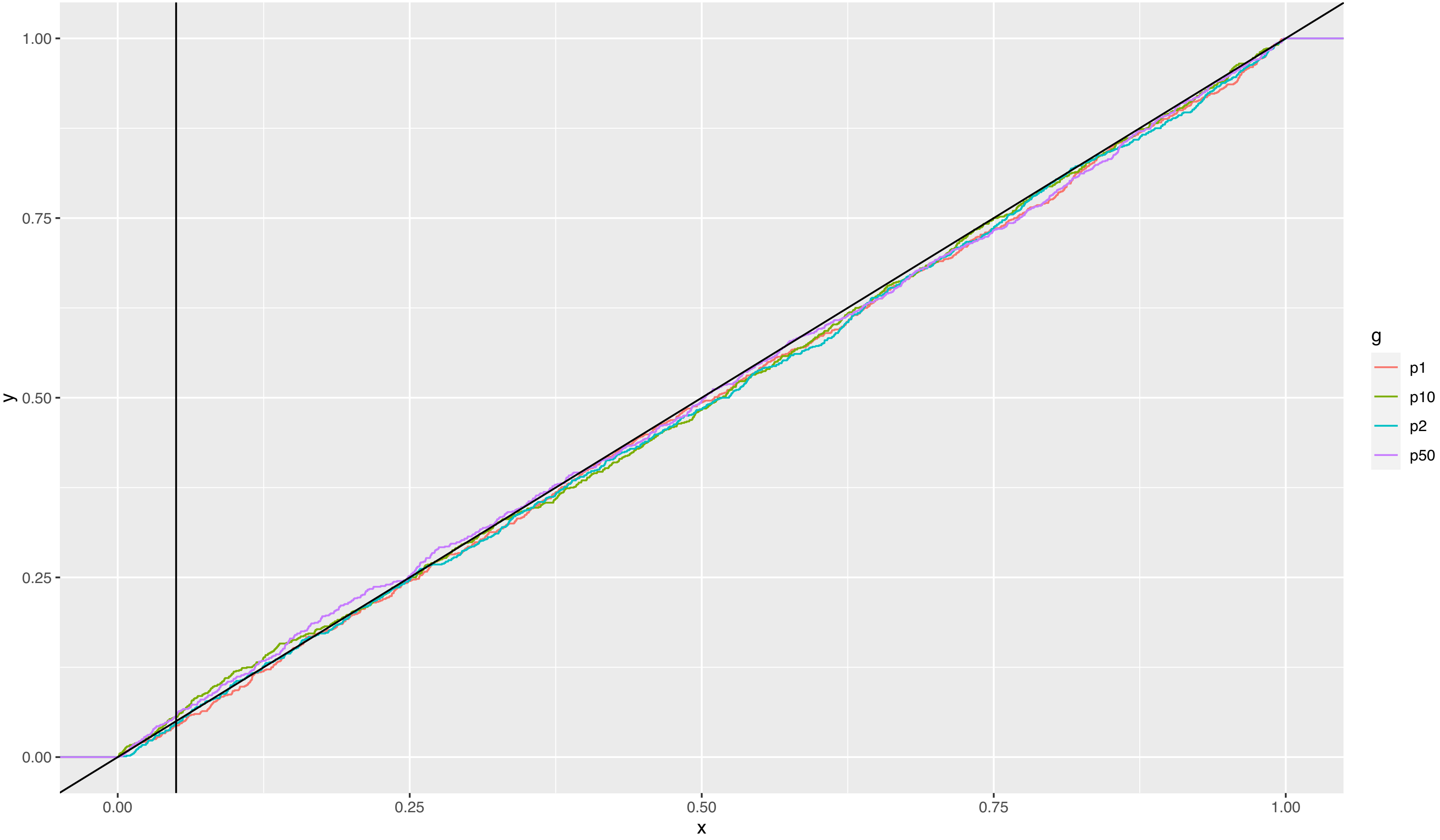
The formulae for these limits are somewhat complicated but easily computable.

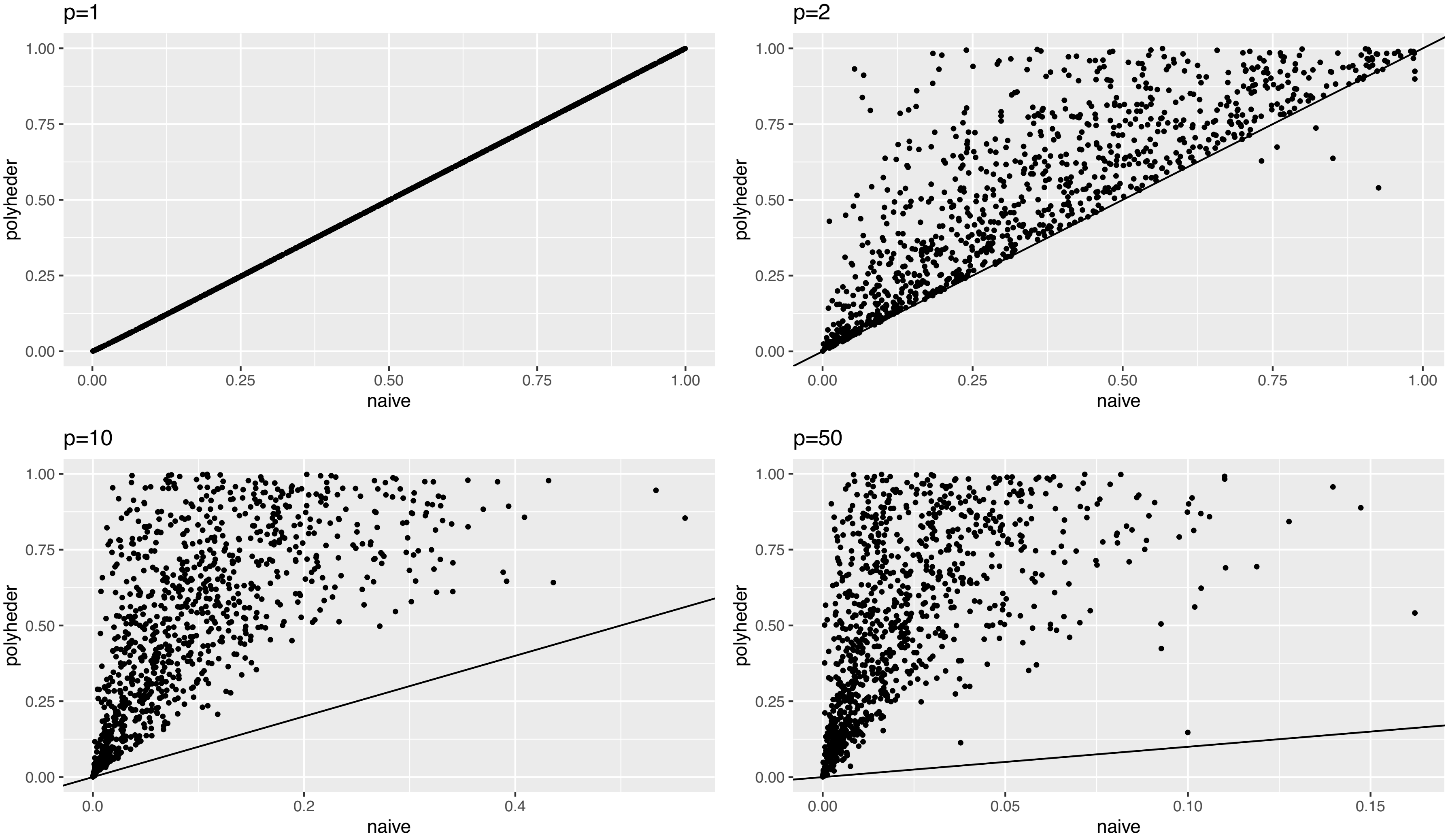
This truncated normal distribution is used to calculate *selection-adjusted* p-values and confidence interval.

(Study Figure 3 in Taylor and Tibshirani (2015).)

also showed figure 6.11 from HTW in class

ECDF of polyheder p-values under the null for first step of forward selection




p=1
p=2
p=10
p=50
polyheder
naive
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
0.0
0.2
0.4
0.00
0.05
0.10
0.15

Polyhedral lasso result

The same methodology can be used for the lasso, here also the selection of predictors can be described as a polyhedral region of the form $Ay \leq b$ - for a fixed value λ.

For the lasso the A and b will depend on

- the predictors
- the active set
- λ

but not on y.

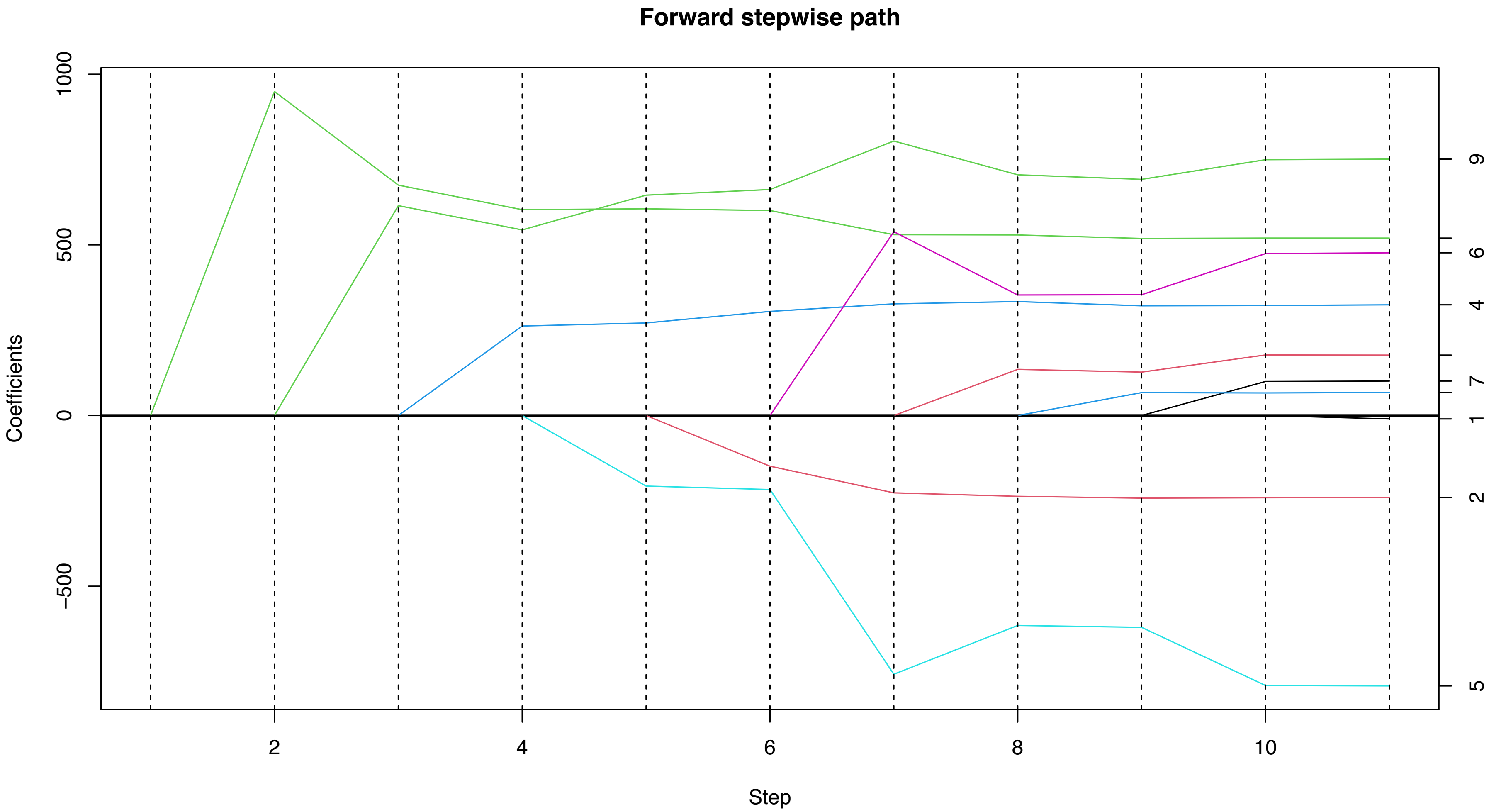
The methods are on closed form, but the values c and d may be of complicated form.

Showed formulas for
polyheder from Taylor
and Tibshirani in class

Selective inference with the diabetes data

Forward selection diabetes

```
      [,1] [,2] 
 [1,] "1"  "age"
 [2,] "2"  "sex"
 [3,] "3"  "bmi"
 [4,] "4"  "map"
 [5,] "5"  "tc" 
 [6,] "6"  "ldl"
 [7,] "7"  "hdl"
 [8,] "8"  "tch"
 [9,] "9"  "ltg"
[10,] "10" "glu"
```


Forward stepwise path
Coefficients
1000
500
0
−500
Step
2
4
6
8
10
9
6
4
7
1
2
5

```
Call:
fsInf(obj = fsfit)

Standard deviation of noise (specified or estimated) sigma

Sequential testing results with alpha = 0.100
 Step Var     Coef Z-score P-value LowConfPt UpConfPt LowTa
    1   3  949.435  17.532   0.000   790.681 1037.113
    2   9  614.951  10.163   0.000   521.696  887.192
    3   4  262.275   4.291   0.010    90.437  363.617
    4   5 -206.670  -3.266   0.684  -279.583 1539.967
    5   2 -148.375  -2.648   0.689  -273.862 1234.380
    6   6  538.586   3.664   0.025   208.452 5364.275
    7   8  135.265   1.121   0.900      -Inf  577.340
    8  10   67.141   1.027   0.033   100.724      Inf
    9   7   99.718   0.470   0.629 -2450.846 1220.006
   10   1  -10.012  -0.168   0.644  -527.324 1058.916

Estimated stopping point from ForwardStop rule = 3
```

For comparison, the suggested forward model with variabls bmi, ltg and map - with naive p-values.

```
Call:
lm(formula = y ~ x[, 3] + x[, 9] + x[, 4])

Residuals:
     Min       1Q   Median       3Q      Max
-140.229  -40.637   -2.187   38.269  139.804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  152.133      2.653  57.342  < 2e-16 ***
x[, 3]       603.074     64.677   9.324  < 2e-16 ***
x[, 9]       543.872     64.619   8.417 5.56e-16 ***
x[, 4]       262.275     62.962   4.166 3.74e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 55.78 on 438 degrees of freedom
Multiple R-squared:  0.4801,    Adjusted R-squared:  0.4765
F-statistic: 134.8 on 3 and 438 DF,  p-value: < 2.2e-16
```

Lasso diabetes

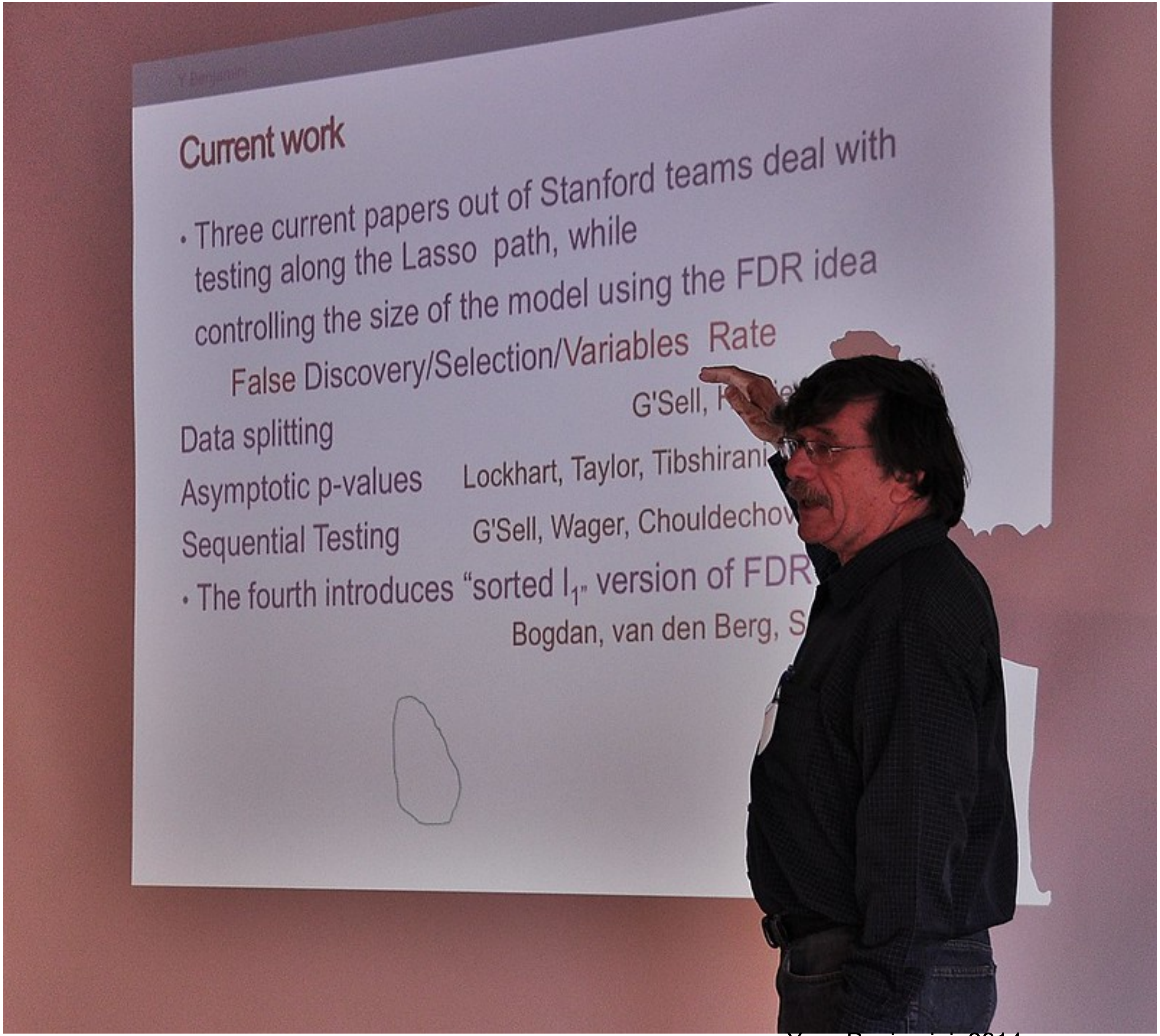
```
[1] 0.2527843
 [1]    0.00000  -33.33808  508.19096  210.35372    0.00000    0.00000
 [7] -138.84778    0.00000  444.56109    0.00000
Call:
fixedLassoInf(x = x, y = y, beta = beta, lambda = lambda * n)

Standard deviation of noise (specified or estimated) sigma = 54.154

Testing results at lambda = 111.731, with alpha = 0.100

 Var     Coef Z-score P-value LowConfPt UpConfPt LowTailArea UpTailArea
   2 -235.776  -3.913   0.117  -325.205   96.516       0.049      0.050
   3  523.562   8.047   0.000   416.203  631.275       0.049      0.049
   4  326.236   5.190   0.000   212.282  430.335       0.048      0.049
   7 -289.117  -4.420   0.003  -397.090 -136.813       0.049      0.050
   9  474.292   7.247   0.000   366.602  582.958       0.050      0.048

Note: coefficients shown are partial regression coefficients
[1] 1.168127e-01 1.092168e-15 3.912618e-05 2.928151e-03 6.562529e-13
```



Y Benjamini

Current work

- Three current papers out of Stanford teams deal with testing along the Lasso path, while controlling the size of the model using the FDR idea

False Discovery/Selection/Variables Rate

Data splitting	G'Sell, K
Asymptotic p-values	Lockhart, Taylor, Tibshirani
Sequential Testing	G'Sell, Wager, Chouldechov

- The fourth introduces "sorted l_1" version of FDR

Bogdan, van den Berg, S

Yoav Benjamini, 2014

Post selection inference and the reproducibility crisis

The *incorrect* use of CIs and p-values in models found from model selection *and* inference on the same data - is though to be one of the main contributors to the *reproducibility crisis in science*. Selective Inference: The Silent Killer of Replicability by Yoav Benjamini Published on Dec 16, 2020