MA8701 Advanced methods in statistical inference and learning

L10: Shrinkage methods for the GLM

↑
logistic regression

Mette Langaas

2/9/23

Lectured 10.02.2023

Before we begin

Literature

- [ELS] The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics, 2009) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Ebook. Chapter 4.4.1-4.4.3 (4.4.4 is covered in 3.2 of HTW).
- [HTW] Hastie, Tibshirani, Wainwright: "Statistical Learning with Sparsity: The Lasso and Generalizations". CRC press. Ebook. Chapter 3.2,3.7, 5.4.3

and for the interested student

- [WNvW] Wessel N. van Wieringen: Lecture notes on ridge regression Chapter 5.

Generalized linear models

(HTW 3.1, 3.2, and TMA4315 GLM background)

The model

The GLM model has three ingredients:

1) Random component
2) Systematic component
3) Link function

We look into that for the ~~normal and~~ binomial distribution - to get multiple linear regression and logistic regression.

- Write in class

1) $Y_i \sim \text{bin}(1, \pi_i)$, $E(Y_i) = \pi_i$

$\underset{N \times 1}{Y}$

2) $\eta_i = x_i^T \beta$ (later $\beta_0 + x_i^T \beta$)

$\underset{N \times p+1}{X}$ $\underset{p+1 \times 1}{\beta}$

3) $\text{logit}(E(Y_i)) = \eta_i$

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \iff \pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

$$(1-\pi_i) = \frac{1+e^{\eta_i} - e^{\eta_i}}{1+e^{\eta_i}} = \frac{1}{1+e^{\eta_i}}$$

$$\pi_i(1-\pi_i) = \frac{e^{\eta_i}}{(1+e^{\eta_i})^2}$$ note this

Q: What is the interpretation of β?

E.g. $\beta = 0$ or $\beta = 1$

Explaining β in logistic regression

- The ratio $\frac{P(Y_i=1)}{P(Y_i=0)} = \frac{\pi_i}{1-\pi_1}$ is called the *odds*.
- If $\pi_i = \frac{1}{2}$ then the odds is 1, and if $\pi_i = \frac{1}{4}$ then the odds is $\frac{1}{3}$. We may make a table for probability vs. odds in R:

pivec	0.10	0.20	0.30	0.40	0.5	0.6	0.70	0.8	0.9
odds	0.11	0.25	0.43	0.67	1.0	1.5	2.33	4.0	9.0

- Odds may be seen to be a better scale than probability to represent chance, and is used in betting. In addition, odds are unbounded above.

We look at the link function (inverse of the response function). Let us assume that our linear predictor has k covariates present

$$
\begin{aligned}
\eta_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \\
\pi_i &= \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \\
\eta_i &= \ln(\frac{\pi_i}{1-\pi_i}) \\
\ln(\frac{\pi_i}{1-\pi_i}) &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \\
\frac{\pi_i}{1-\pi_i} &= \frac{P(Y_i = 1)}{P(Y_i = 0)} = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdots \exp(\beta_k x_{ik})
\end{aligned}
$$

We have a *multiplicative model* for the odds.

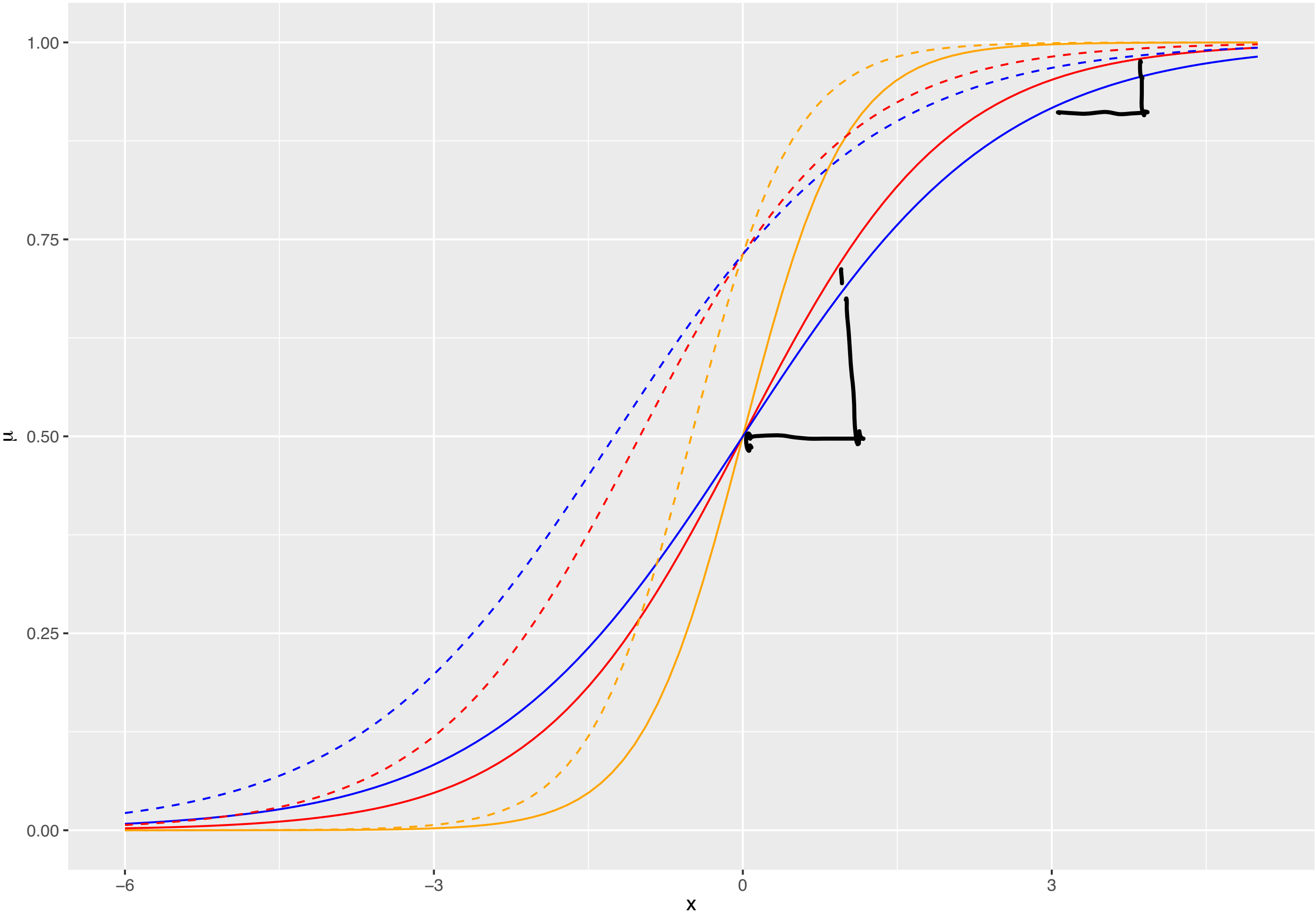
So, what if we increase x_{1i} to $x_{1i}+1$?

If the covariate x_{1i} increases by one unit (while all other covariates are kept fixed) then the odds is multiplied by $\exp(\beta_1)$:

$$
\begin{aligned}
\frac{P(Y_i = 1 \mid x_{i1}+1)}{P(Y_i = 0) \mid x_{i1}+1)} &= \exp(\beta_0)\cdot\exp(\beta_1(x_{i1}+1))\cdots\exp(\beta_k x_{ik}) \\
&= \exp(\beta_0)\cdot\exp(\beta_1 x_{i1})\exp(\beta_1)\cdots\exp(\beta_k x_{ik}) \\
&= \frac{P(Y_i = 1 \mid x_{i1})}{P(Y_i = 0 \mid x_{i1})}\cdot\exp(\beta_1)
\end{aligned}
$$

This means that if x_{i1} increases by 1 then: if $\beta_1 < 0$ we get a decrease in the odds, if $\beta_1 = 0$ no change, and if $\beta_1 > 0$ we have an increase. In the logit model $\exp(\beta_1)$ is easier to interpret than β_1.

The response function as a function of the covariate x and not of η. Solid lines: $\beta_0 = 0$ and β_1 is 0.8 (blue), 1 (red) and 2 (orange), and dashed lines with $\beta_0 = 1$.



Parameter estimation

First logistic regression, then ridge and lasso logistic regression - and (maybe) elastic net logistic regression.

Logistic regression

- Maximum likelihood estimation $=$ maximize the likelihood of the data. We write for the loglikelihood $l(\beta_0, \beta; y, X)$.
- We write out the loglikelihood for the binomial with logit link $=$logistic regression.

$$L(\beta) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

if no continuous covariates we may have i to be a covariate pattern of n_i obs $\binom{n_i}{y_i}$

$$l(\beta) = \ln L(\beta) = \sum_{i=1}^{N} y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)$$

$$= \sum_{i=1}^{N} y_i \underbrace{(\ln \pi_i - \ln(1-\pi_i))}_{\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = x_i^T\beta} + \underbrace{\ln(1-\pi_i)}_{-\ln(1+e^{x_i^T\beta})}$$

$$= \sum_{i=1}^{N}\left(y_i x_i^T\beta - \ln(1+e^{x_i^T\beta})\right)$$

concave loglikhood

(all ok except is separable problem)

score equation $\dfrac{\partial l}{\partial \beta} = \underset{(p+1)\times 1}{0}$

$$\frac{\partial l}{\partial \beta} = \sum_{i=1}^{N} y_i \overset{\overset{p+1\times 1}{\downarrow}}{x_i} - \underbrace{\frac{1}{1+e^{x_i^T\beta}} e^{x_i^T\beta}}_{\pi_i} \cdot \overset{\overset{p+1\times 1}{|}}{x_i}$$

$$\frac{\partial\, d^T\beta}{\partial \beta} = d$$

$$\frac{\partial\, \beta^T A \beta}{\partial \beta} = 2A\beta \quad \text{if } A = A^T$$

$$= \sum_{i=1}^{N} x_i (y_i - \pi_i) \underset{\uparrow}{= 0}$$

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

(p+1) nonlin eq's

observe first element:

$$\sum_{i=1}^{N} 1 \cdot (y_i - \pi_i) = 0 \Leftrightarrow \underset{\substack{\uparrow \\ \text{exp. \# cases}}}{\sum_{i=1}^{N} \pi_i} = \underset{\substack{\uparrow \\ \text{observed \# cas}}}{\sum_{i=1}^{N} y_i}$$

Algorithms

To understand the ridge and lasso logistic regression we first look at the *iteratively reweighted least squares* (IRLS) - as a result of the Newton Raphson method for the logistic regression (unpenalized).

Properties

The parameter estimator is asymptotically normal. Unbiased with variance the inverse of the Fisher information matrix - as known TMA4315.

$f(x)=0$

univariate

$$f(x) \approx f(x_0) + (x-x_0) \frac{\partial f}{\partial x}\Big|_{x=x_0}$$

$$x = x_0 - \left(\frac{df}{dx}\Big|_{x=x_0}\right)^{-1} f(x_0) \quad \circlearrowleft$$

multivariate:

$$\vec{f}(x) \approx \vec{f}(x_0) + J\vec{f}\,\Big|_{x=x_0} (x-x_0)$$

Our eq is $\frac{\partial l}{\partial \beta} = 0$; $\frac{\partial l}{\partial \beta}\Big|_{\beta^{old}} + (\beta^{new} - \beta^{old}) \frac{\partial^2 l}{\partial\beta\partial\beta^T}\Big|_{\beta^{old}} = 0$

β^{old}

$$\underset{(p+1 \times 1)}{\beta^{new}} = \beta^{old} - \underset{p+1 \times p+1}{\left(\frac{\partial^2 l}{\partial\beta\partial\beta^T}\right)^{-1}} \underset{p+1 \times 1}{\frac{\partial^2 l(\beta)}{\partial \beta}}$$

$\Rightarrow$ need $\frac{\partial^2 l}{\partial\beta\partial\beta^T}$

$$\frac{\partial l}{\partial\beta} = \sum_{l=1}^{N} x_i(y_i - \pi_i) \qquad \pi_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

$$\frac{\partial l}{\partial\beta} = \overset{p+1 \times N}{X^T}(\overset{N \times 1}{Y} - \overset{N \times 1}{\pi})$$

$$\frac{\partial^2 l}{\partial\beta\partial\beta^T} = 0 - \frac{\partial}{\partial\beta^T}\left(\sum x_i \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)$$

$x_i e^{x_i^T\beta} . (1+e^{x}$

$$= -\sum_{i=1}^{N} x_i \frac{e^{x_i^T\beta}\cdot x_i^T\cdot(1+e^{x_i^T\beta}) - e^{x_i^T\beta}\cdot x_i^T e^{x_i^T\beta}}{(1+e^{x_i^T\beta})^2}$$

$$= -\sum_{i=1}^{N} x_i x_i^T \frac{e^{x_i^T\beta}}{(1+e^{x_i^T\beta})^2}\left[\underbrace{(1+e^{x_i^T\beta}) - e^{x_i^T\beta}}_{1}\right]$$

$$= -\sum_{i=1}^{N} x_i x_i^T \;\pi_i(1-\pi_i)$$

$$\frac{\partial^2 l}{\partial\beta\,\partial\beta^T} = -X^T W X$$

$$W = \mathrm{diag}(\pi_i(1-\pi_i))$$

$$H = -\frac{\partial^2 l}{\partial\beta\,\partial\beta^T}$$

Fisherinfo = $E(H)$

Newton-Raphson:

Why not Fisher Scoring? → $H = E(H)$ GLM canonical link

$$\beta^{new} = \beta^{old} - \overset{E}{\left(\frac{\partial^2 l}{\partial\beta\,\partial\beta^T}\right)^{-1}} \frac{\partial l(\beta)}{\partial\beta}$$

$$\underset{\overset{\uparrow}{I}}{=}\beta^{old} + X^T W^{old} X \cdot X^T \underset{\overset{\uparrow}{I}}{(} Y - \pi^{old})$$

$\beta \to \pi \to W$

weighted LS

min $(Y - X\beta^T)W(Y - X\beta)$

$\hat{\beta}_{WLS} = (X^TWX)^{-1}X^TWY$ → WLS form

$$= \overbrace{(X^T W^{old} X)^{-1} X^T W^{old} X}^{I} \beta^{old}$$

$$+ X^T W^{old} X \cdot X^T \underbrace{W^{old} W^{old^{-1}}}_{I} (Y - \pi^{old})$$

$$= (X^T W^{old} X)^{-1} X^T W^{old} \underbrace{\left(X\beta^{old} + (W^{old})^{-1}(Y - \pi^{old})\right)}_{z^{old}}$$

adjusted response

$$\beta^{new} = (X^T W^{old} X)^{-1} X^T W^{old} Z^{old}$$

NB

→ we know this is the solution to $\underset{\beta}{\operatorname{argmin}} \left\{ (Z^{old} - X\beta^{old})^T W^{old} (Z^{old} - X\beta^{old}) \right\}$

Iterate until convergence this IRWLS iterated reweighted least sq

In class we now scroll down to the South African data set and look at the data and the logistic regression.

Example: South African heart disease

(ELS 4.4.2)

Group discussion

Comment on what is done and the results. Where are the CIs and p-values for the ridge and lasso version?

Data set

The data is presented in ELS Section 4.4.2, and downloaded from http://statweb.stanford.edu/~tibs/ElemStatLearn.1stEd/ with information in the file `SAheat.info` and data in `SAheart.data`.

- This is a retrospective sample of males in a heart-disease high-risk region in South Africa.
- It consists of 462 observations on the 10 variables. All subjects are male in the age range 15-64.
- There are 160 cases (individuals who have suffered from a conorary heart disease) and 302 controls (individuals who have not suffered from a conorary heart disease).
- The overall prevalence in the region was 5.1%.

The response value (chd) and covariates

- chd : conorary heart disease {yes, no} coded by the numbers {1, 0}
- sbp : systolic blood pressure
- tobacco : cumulative tobacco (kg)
- ldl : low density lipoprotein cholesterol
- adiposity : a numeric vector
- famhist : family history of heart disease. Categorical variable with two levels: {Absent, Present}.
- typea : type-A behavior
- obesity : a numerical value
- alcohol : current alcohol consumption
- age : age at onset

The goal is to identify important risk factors.

||

model selection or just sign. effects

Data description

We start by loading and looking at the data:

```
ds=read.csv("./SAheart.data",sep=",")[,-1]
ds$chd=as.factor(ds$chd)
ds$famhist=as.factor(ds$famhist)
dim(ds)
```

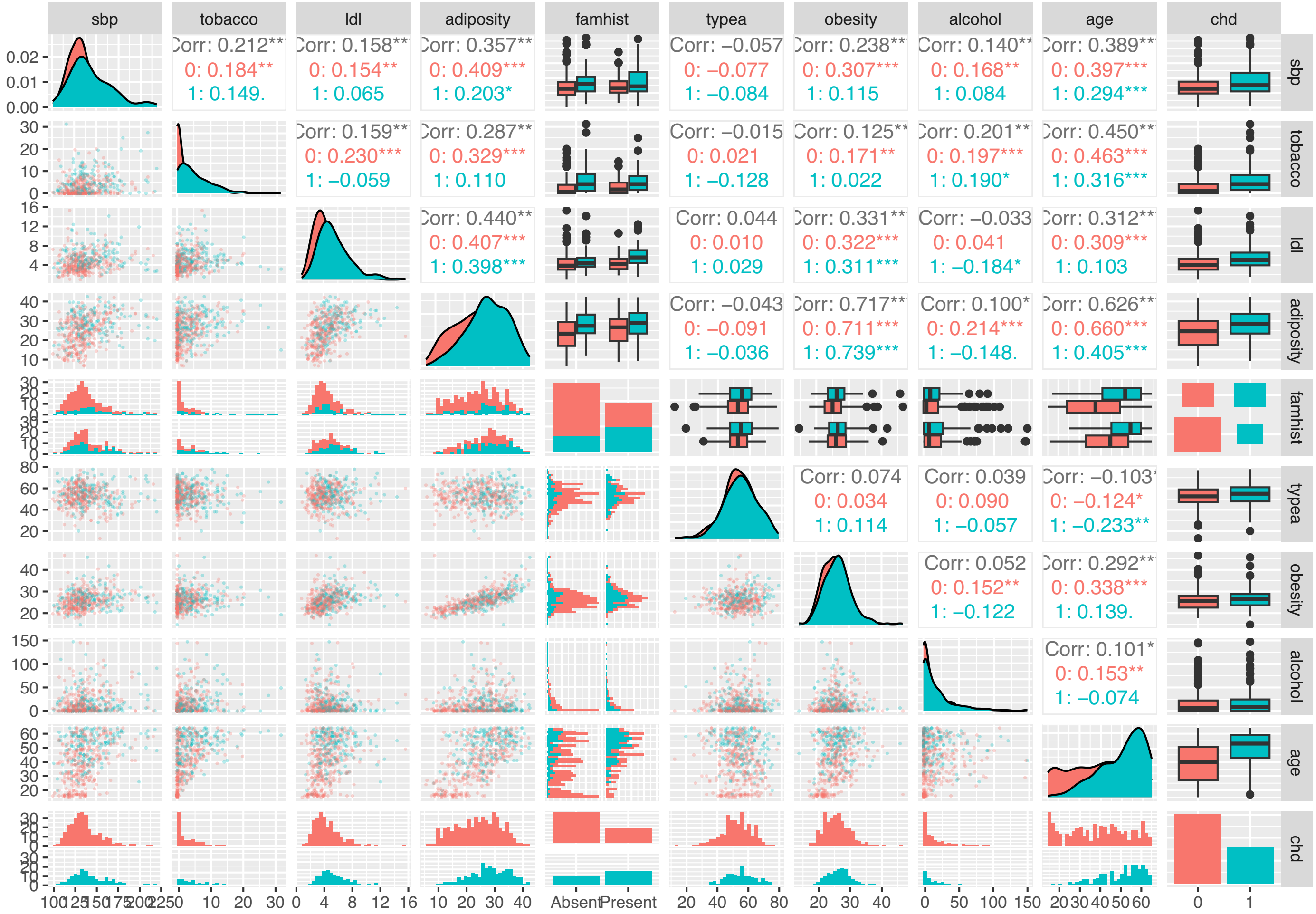
```
[1] 462  10
```

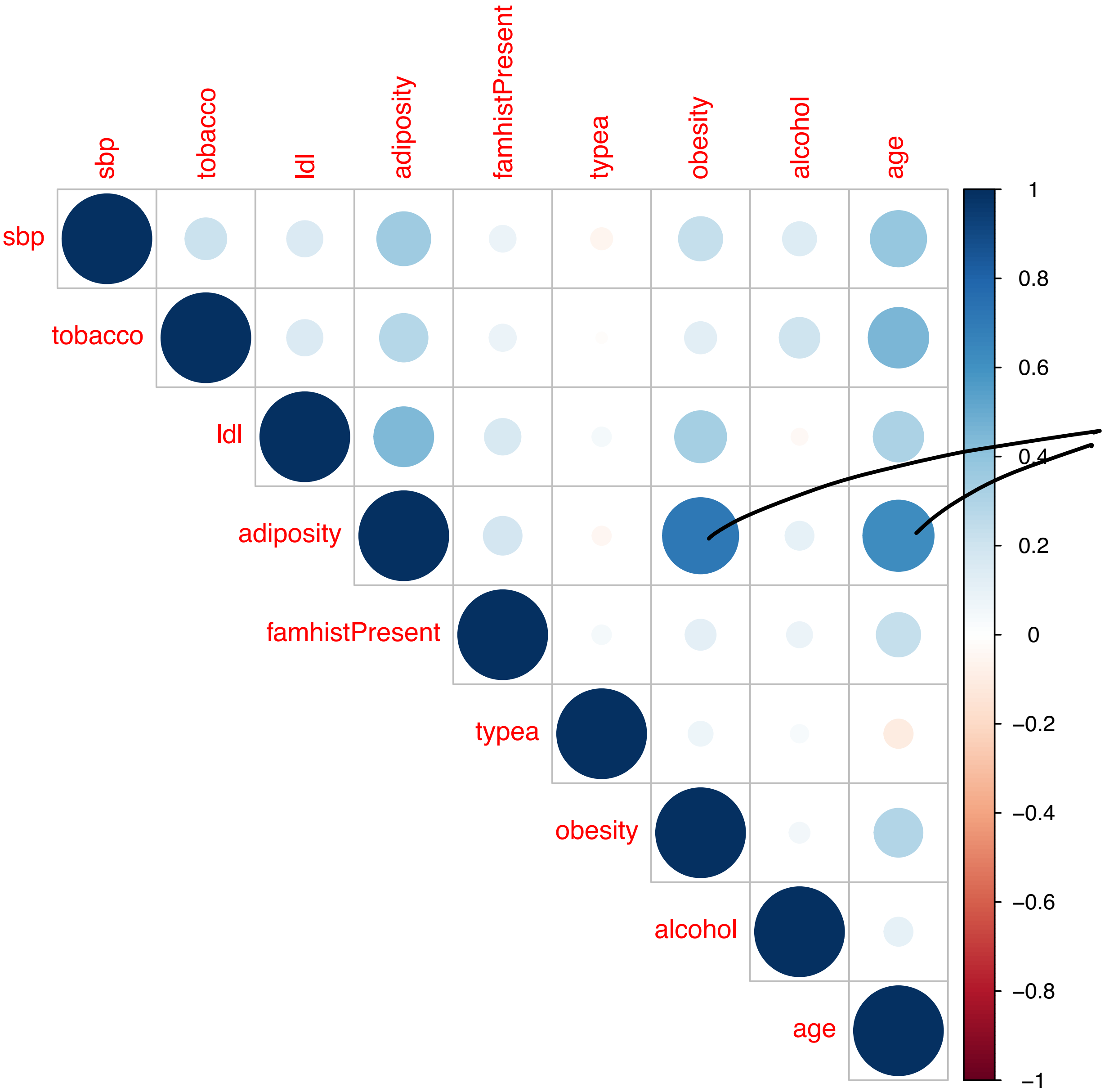
```
colnames(ds)
```

```
 [1] "sbp"        "tobacco"    "ldl"        "adiposity" "famhi
 [7] "obesity"    "alcohol"    "age"        "chd"
```

```
head(ds)
```

```
  sbp tobacco  ldl adiposity famhist typea obesity alcohol
1 160   12.00 5.73     23.11 Present    49   25.30   97.20
2 144    0.01 4.41     28.61  Absent    55   28.87    2.06
3 118    0.08 3.48     32.28 Present    52   29.14    3.81
4 170    7.50 6.41     38.03 Present    51   31.99   24.26
5 134   13.60 3.50     27.78 Present    60   25.99   57.34
6 132    6.20 6.47     36.21 Present    62   30.77   14.14
```


sbp
tobacco
ldl
adiposity
famhist
typea
obesity
alcohol
age
chd
Corr: 0.212**
0: 0.184**
1: 0.149.
Corr: 0.158**
0: 0.154**
1: 0.065
Corr: 0.357**
0: 0.409***
1: 0.203*
Corr: −0.057
0: −0.077
1: −0.084
Corr: 0.238**
0: 0.307***
1: 0.115
Corr: 0.140**
0: 0.168**
1: 0.084
Corr: 0.389**
0: 0.397***
1: 0.294***
Corr: 0.159**
0: 0.230***
1: −0.059
Corr: 0.287**
0: 0.329***
1: 0.110
Corr: −0.015
0: 0.021
1: −0.128
Corr: 0.125**
0: 0.171**
1: 0.022
Corr: 0.201**
0: 0.197***
1: 0.190*
Corr: 0.450**
0: 0.463***
1: 0.316***
Corr: 0.440**
0: 0.407***
1: 0.398***
Corr: 0.044
0: 0.010
1: 0.029
Corr: 0.331**
0: 0.322***
1: 0.311***
Corr: −0.033
0: 0.041
1: −0.184*
Corr: 0.312**
0: 0.309***
1: 0.103
Corr: −0.043
0: −0.091
1: −0.036
Corr: 0.717**
0: 0.711***
1: 0.739***
Corr: 0.100*
0: 0.214***
1: −0.148.
Corr: 0.626**
0: 0.660***
1: 0.405***
Corr: 0.074
0: 0.034
1: 0.114
Corr: 0.039
0: 0.090
1: −0.057
Corr: −0.103
0: −0.124*
1: −0.233**
Corr: 0.052
0: 0.152**
1: −0.122
Corr: 0.292**
0: 0.338***
1: 0.139.
Corr: 0.101*
0: 0.153**
1: −0.074
0.02
0.01
0.00
30
20
10
0
16
12
8
4
40
30
20
10
80
60
40
20
150
100
50
0
60
50
40
30
20
100 125 150 175 200 225
0 10 20 30
0 4 8 12 16
10 20 30 40
Absent Present
20 40 60 80
20 30 40
0 50 100 150
20 30 40 50 60
0 1



may lead to models difficult to explain

Logistic regression

We now fit a (multiple) logistic regression model using the `glm` function and the full data set. In order to fit a logistic model, the `family` argument must be set equal to `="binomial"`. The `summary` function prints out the estimates of the coefficients, their standard errors and z-values. As for a linear regression model, the significant coefficients are indicated by stars where the significant codes are included in the `R` printout.

```r
glm_heart = glm(chd~.,data=dss, family="binomial")
summary(glm_heart)
```

```

Call:
glm(formula = chd ~ ., family = "binomial", data = dss)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7781  -0.8213  -0.4387   0.8889   2.5435

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.878545   0.123218  -7.130  1.0e-12 ***
sbp             0.133308   0.117452   1.135 0.256374
tobacco         0.364578   0.122187   2.984 0.002847 **
ldl             0.360181   0.123554   2.915 0.003555 **
adiposity       0.144616   0.227892   0.635 0.525700
famhistPresent  0.456538   0.112433   4.061  4.9e-05 ***
typea           0.388726   0.120954   3.214 0.001310 **
```

A very surprising result here is that `sbp` and `obesity` are NOT significant and `obesity` has negative sign. This is a result of the correlation between covariates. In separate models with only `sbp` or only `obesity` each is positive and significant.

Q: How would you interpret the estimated coefficient for `tobacco`?

Penalized logistic regression

- For penalized method we instead minimize the negative loglikelihood scaled with $\frac{1}{N}$.
- The ridge and lasso penalty is added to the scaled negative loglikelihood.
- Write in class

Penalized regression

$$\max_{\beta_0,\beta}\left\{ \frac{1}{N} l(\beta) - \lambda \begin{cases} \sum^{p} \beta_j^2 \\ \sum_{j=1}^{p} |\beta_j| \end{cases} \right\}$$

$$\lambda \sum^{p} \beta_j^2$$

$$\min_{\beta_0,\beta_1} -\frac{1}{N} \sum_{i=1}^{N} \left\{ y_i x_i^T \beta + \ln\left(1+e^{x_i^T \beta}\right) \right\} + \lambda \sum_{j=1}^{p} |\beta_j|$$

Remark: $Y \in \{0,1\}$

machine learning $\{-1,1\} \rightarrow \left[\frac{1}{N} \sum_{i=1}^{N} \ln\left(1+e^{\widetilde{-y_i x}^T \beta}\right) \right]$

RIDGE LOGISTIC

→ Add score & Hessian from penalization term

$$l_{pen}(\beta) = l(\beta) - \lambda \beta^T \beta$$

$$\frac{\partial l_{pen}}{\partial \beta} = \frac{\partial l}{\partial \beta} - \lambda \beta \qquad \frac{\partial^2 l_{pen}}{\partial \beta \partial \beta^T} = \frac{\partial^2 l}{\partial \beta \partial \beta^T} - \lambda I$$

[penalize the intercept? usually NOT, just proceed - fix later]

Suggestion for

$$\frac{\partial l_{pen}}{\partial \beta} = \frac{\partial l}{\partial \beta} - \begin{bmatrix} 0 \\ \lambda \beta \end{bmatrix} \qquad \underset{p+1 \times p+1}{\frac{\partial^2 l_{pen}}{\partial \beta \partial \beta^T}} = \overset{-X^TWX}{\frac{\partial^2 l}{\partial \beta \partial \beta^T}} - \lambda \overset{p \times p}{I_{pp}}$$

$$\begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \lambda I_{pp} & \\ 0 & & \end{bmatrix}$$

↓ Here intercept also penalized

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l_{pen}}{\partial\beta\,\partial\beta^T}\right)^{-1}\frac{\partial l_{pen}}{\partial\beta}$$

$$= \beta^{old} + \left(X^T \overset{old}{W} X + \lambda I\right)^{-1}\left[X^T(y-\pi^{old}) - \lambda\beta^{old}\right]$$

$$= \cdots = \left(X^T W^{old} X + \lambda I\right)^{-1} X^T W^{old} z^{old}$$

where $W^{old} = \text{diag}\left(\pi_i^{old}(1-\pi_i^{old})\right)$

$$z^{old} = X\beta^{old} + W^{old\,-1}(y - \pi^{old})$$

🟩 $\beta + V^{-1}(X^T(y-\pi) - \lambda\beta)$

$= V^{-1}V\beta - \lambda V^{-1}\beta + V^{-1}X\,WW^{-1}(y-\pi)$

$= V^{-1}X^TW\underbrace{\left(X\beta + W^{-1}(y-\pi)\right)}_{z}$

just check of 🟩

$V^{-1}\left(\underbrace{\overset{X^TWX+\lambda I}{V}\beta - \lambda\beta}\right) =$

$= V^{-1}\left((X^TWX+\cancel{\lambda} I)\beta - \cancel{\lambda\beta}\right) = V^{-1}X^TWX\beta$

What if $\lambda \to \infty \Rightarrow \beta \to 0$

If β_0 unpenalized then even if the other β's $\to 0$ then
β_0 will model the success prob. see WNvW 5.2

Algorithms

- The likelihood for the GLM is differentiable, and the ridge and lasso objective functions are convex - and can be solved with socalled "standard convex optimization methods".
- But, by popular demand also special algorithms are available - building on the cyclic coordinate descent.

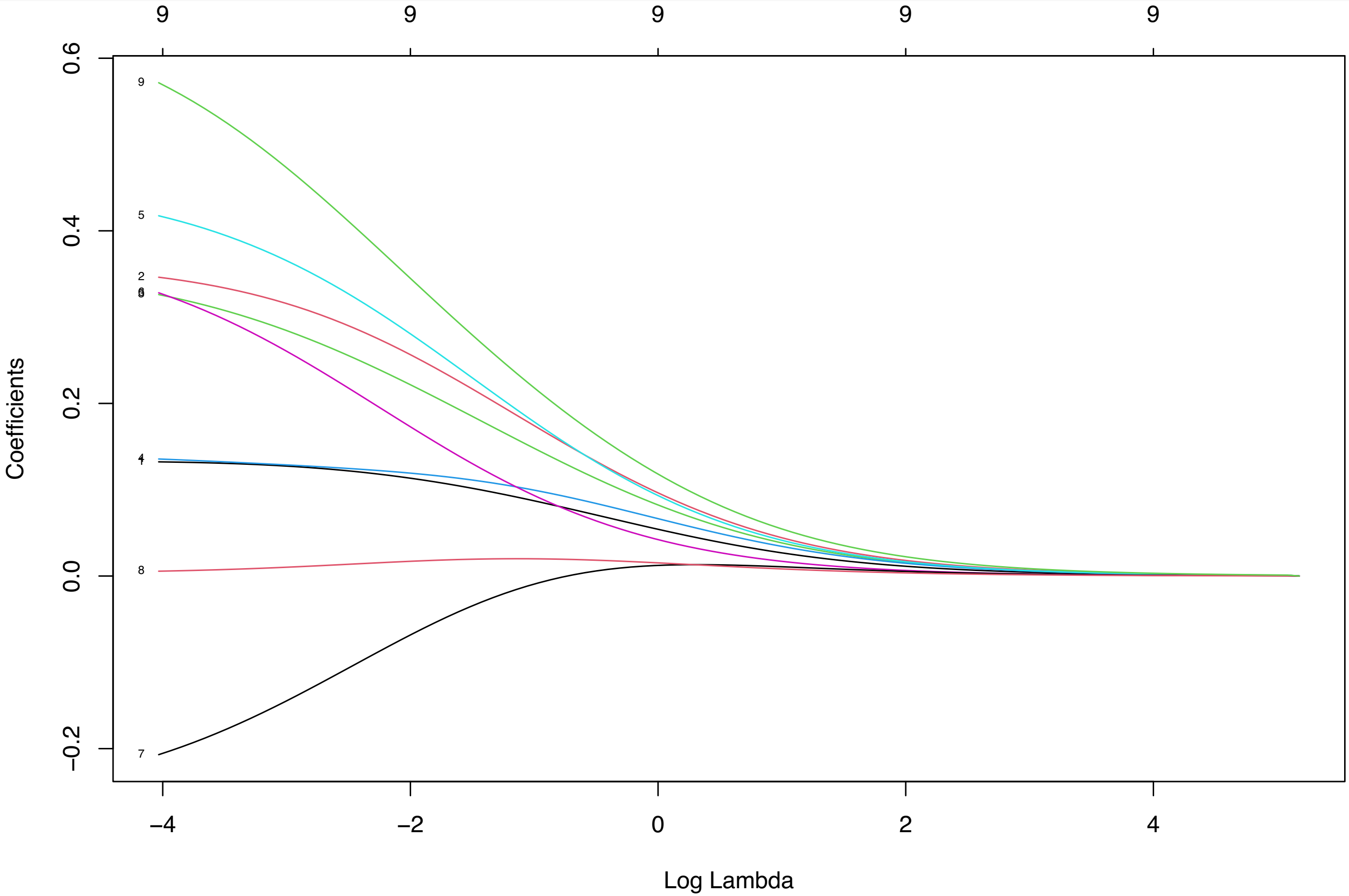
Ridge logistic regression IRWLS

SA dataset → how to choose the λ?

Default deviance used for choos λ
+ 10 fold CV

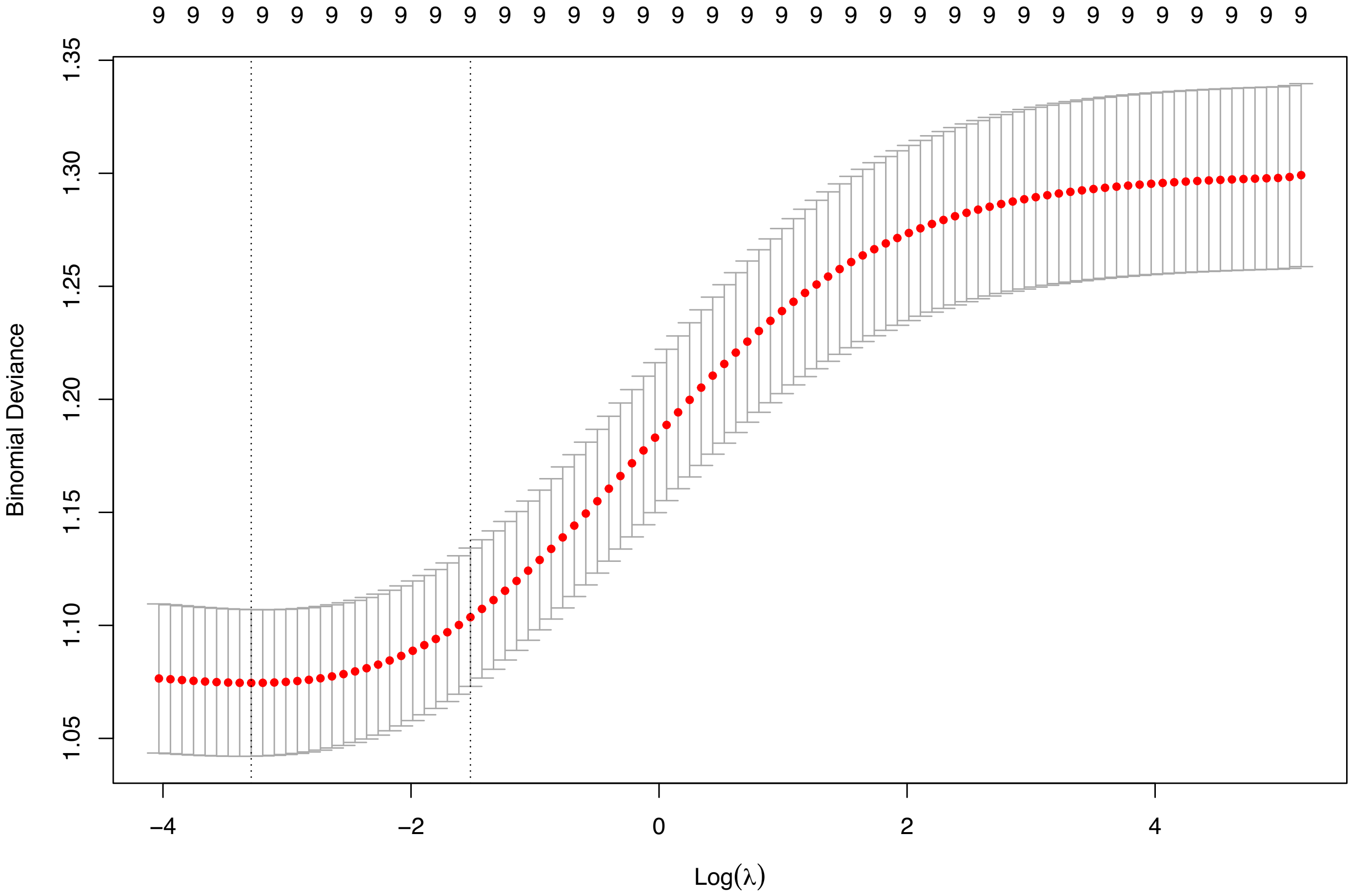
Ridge logistic regression

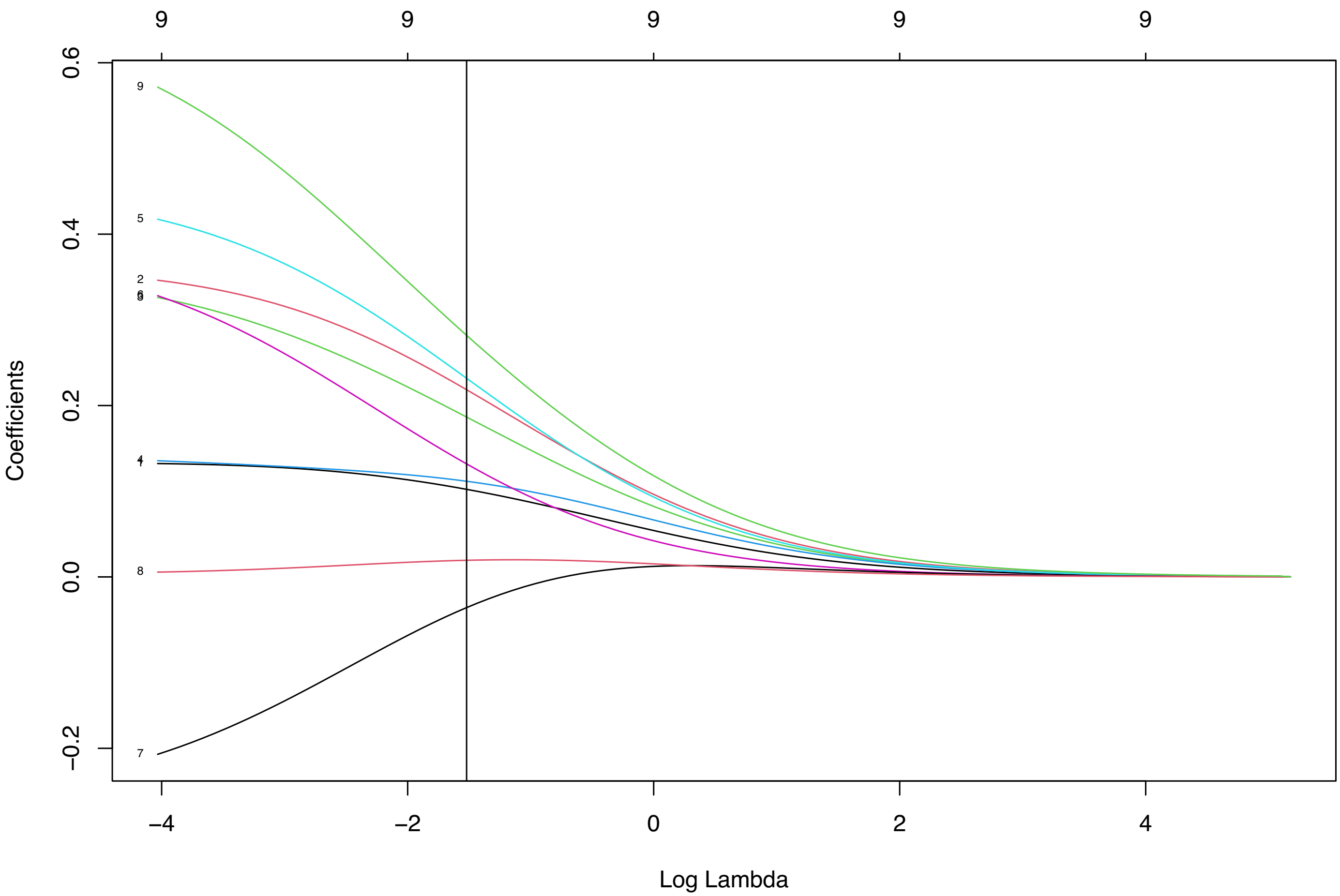
```
ridgefit=glmnet(x=xss,y=ys,alpha=0,standardize=FALSE,family
plot(ridgefit,xvar="lambda",label=TRUE)
```



```
[1] "The lamda giving the smallest CV error 0.0373130776476

[1] "The 1sd err method lambda 0.218543472568106"
```




9
9
9
9
9
0.6
0.4
0.2
0.0
−0.2
Coefficients
9
5
2
6
3
4
1
8
7
−4
−2
0
2
4
Log Lambda

Deviance

The *deviance* is based on the likelihood ratio test statistic.
The derivation assumes that data can be grouped into covariate patterns, with G groups (else interval solutions are used in practice).
Saturated model: If we were to provide a perfect fit to our data then we would estimate π_j by the observed frequency for the group, $\hat{y}_j = y_j$.
Candidate model: the model with the current choice of λ.

$$D_\lambda = 2(l(\text{saturated model}) - l(\text{candidate model}_\lambda))$$

The **null deviance** is replacing the candidate model with a model where $\hat{y}_i = \frac{1}{N}\sum_{i=1}^{N} y_i$ (the case proportion).

Criteria for choosing λ

We use cross-validation to choose λ.

For regression we choose λ by minimizing the (mean) squared error.

For (ridge and) lasso logistic regression we may choose:

- misclassification error rate on the validation set
- ROC-AUC or PR-AUC
- binomial deviance

intercept=TRUE
in glmnet

logistic

ridge

```
10 x 2 sparse Matrix of class "dgCMatrix"
                         s1
(Intercept)    -0.71220689 -0.878545196
sbp             0.10221203  0.133308398
tobacco         0.21846208  0.364577926
ldl             0.18656817  0.360180594
adiposity       0.11163533  0.144616485
famhistPresent  0.23181050  0.456537713
typea           0.13189202  0.388725509
obesity        -0.03579032 -0.265082072
alcohol         0.01941844  0.002978424
age             0.28192570  0.660695163
```

LASSO LOGISTIC

Remember for ord logistic: at each step in the N-R

$$\underset{\beta_0,\beta}{\text{minimize}} \; (z^{old} - X\beta)^T W^{old} (z^{old} - X\beta)$$

↖ can be seen as a quad approx to neg log likelihood

What if we replace this by

$$\underset{\beta_0,\beta}{\text{minim}} \; (z^{old} - X\beta)^T W^{old} (z^{old} - X\beta) + \lambda \sum_{j=1}^{p} |\beta_j|$$

extra inner loop

We know how to solve this (L8) by cyclic coord descent

except we now have $\overset{\text{diag}}{W}$ to take into account

by regarding z^{old} and w^{old} to be constants loop over j and work with partial residuals

$$\hat{\beta}_{lasso,j} = \text{sign}(\hat{\beta}_{WLS,j})\left(|\hat{\beta}_{WLS,j}| - \frac{\lambda}{2}\right)_+$$

where $\hat{\beta}_{WLS} = (X^TWX)^{-1}X^TWZ$

If elastic net linear model $\Rightarrow$ $$\hat{\beta}_{el,j} = \frac{1}{\underset{w}{\sum_{i=1}^{N}} x_{ij}^2 + \lambda(1-\alpha)} S_{\lambda\alpha}\left(\underset{w}{\sum_{i=1}^{N}} r_{ij} x_{ij}\right)$$

$$\underset{w}{r_{ij}} = y_i - \hat{\beta}_0 - \sum x_{ik}\hat{\beta}_k$$

plus add
and replace y by z
in r

Lasso logistic regression fitting algoritm

(HTW page 116)

```
OUTER LOOP: start with lambdamax and decrement              $\beta$     $\beta$

                                     $z = X\beta + W^{-1}(y - \pi)$
    MIDDLE LOOP (with warm start)

                      $(z - X\beta)^T W (z - X\beta)$
        compute quadratic approximation
        for current beta-estimates

            INNER LOOP: cyclic coordinate descent
            to minimize quadratic approximation
            added the lasso penalty
```

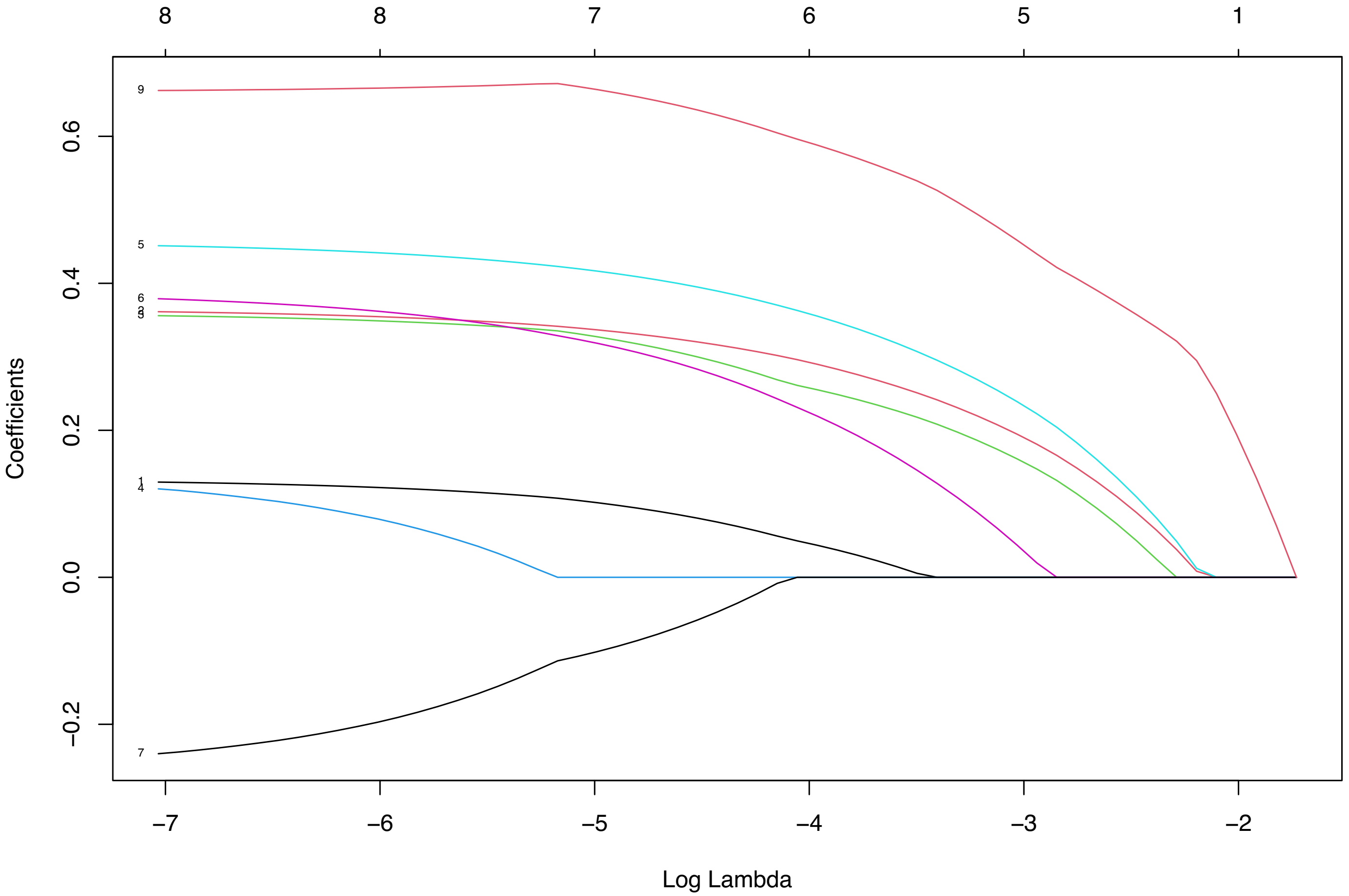
Lasso logistic regression

Numbering in plots is order of covariates, so:

```
cbind(1:9,colnames(xss))
```

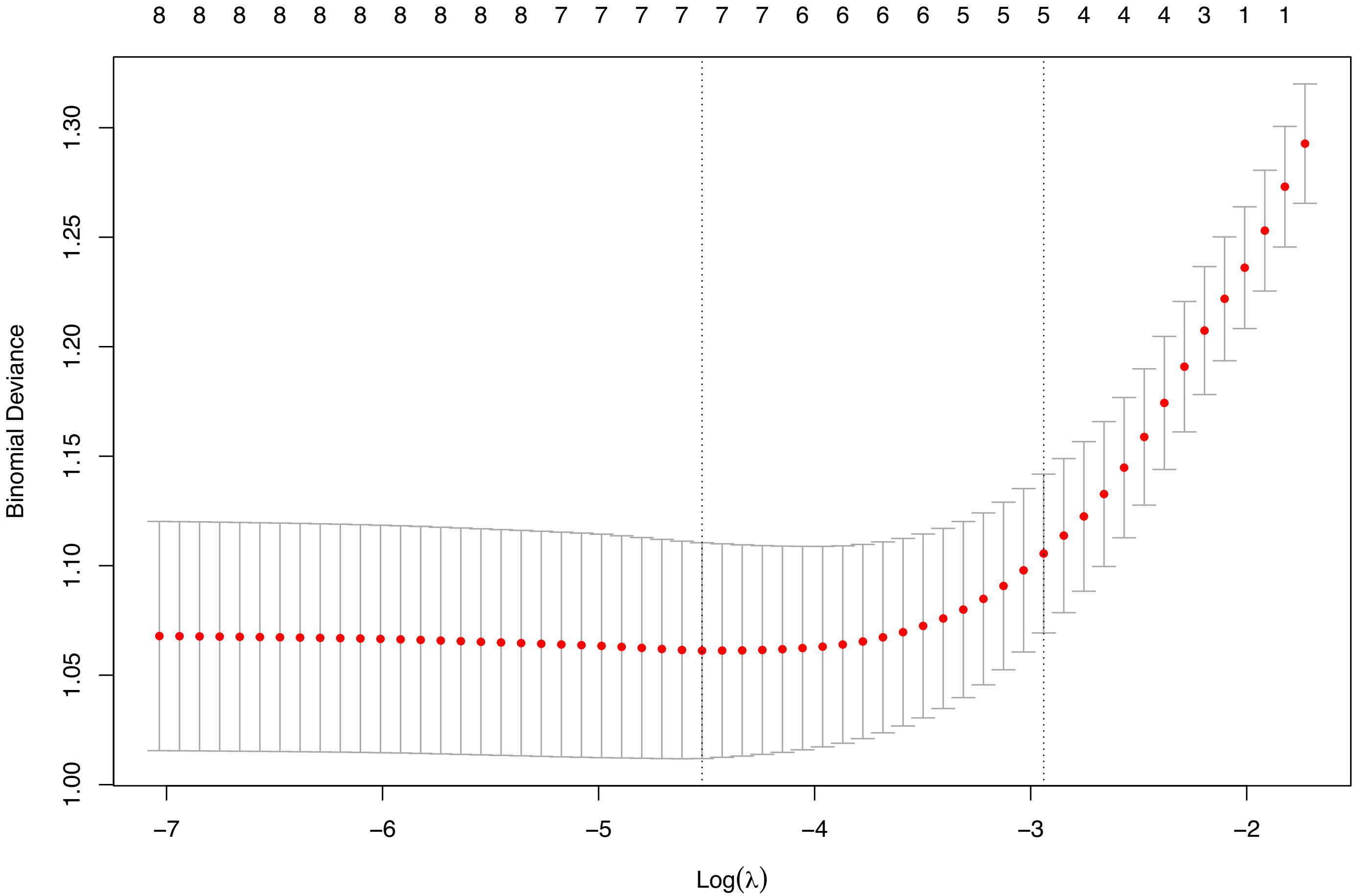
```
      [,1] [,2]
 [1,] "1"  "sbp"
 [2,] "2"  "tobacco"
 [3,] "3"  "ldl"
 [4,] "4"  "adiposity"
 [5,] "5"  "famhistPresent"
 [6,] "6"  "typea"
 [7,] "7"  "obesity"
 [8,] "8"  "alcohol"
 [9,] "9"  "age"
```

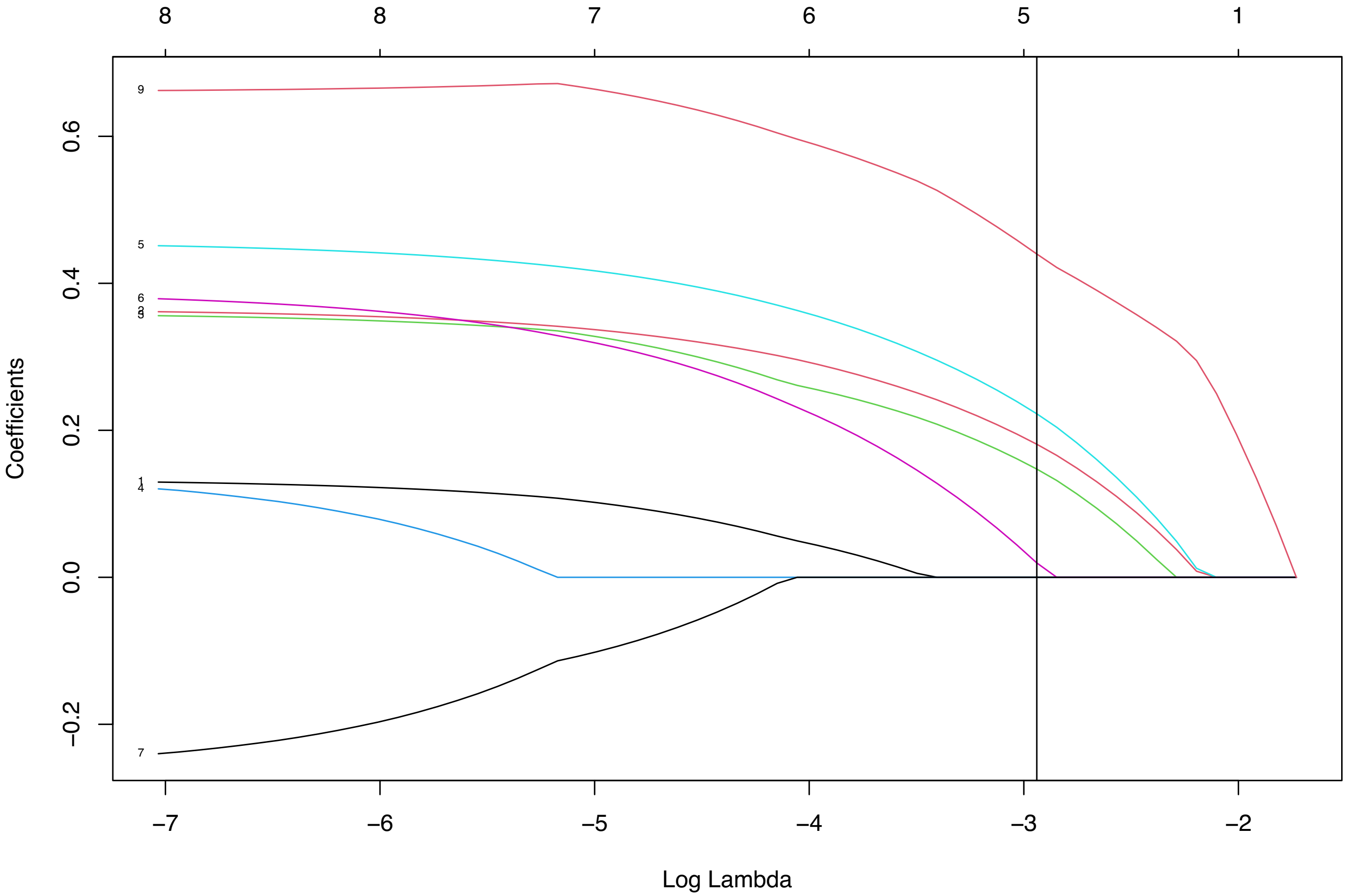
```
lassofit=glmnet(x=xss,y=ys,alpha=1,standardize=FALSE,famil
```


8
8
7
6
5
1
9
5
6
2
3
1
4
7
0.6
0.4
0.2
0.0
−0.2
Coefficients
−7
−6
−5
−4
−3
−2
Log Lambda

```
[1] "The lamda giving the smallest CV error 0.0108769601280

[1] "The 1sd err method lambda 0.052890323504839"
```




8
8
7
6
5
1
9
5
6
2
3
1
4
7
0.6
0.4
0.2
0.0
−0.2
Coefficients
−7
−6
−5
−4
−3
−2
Log Lambda

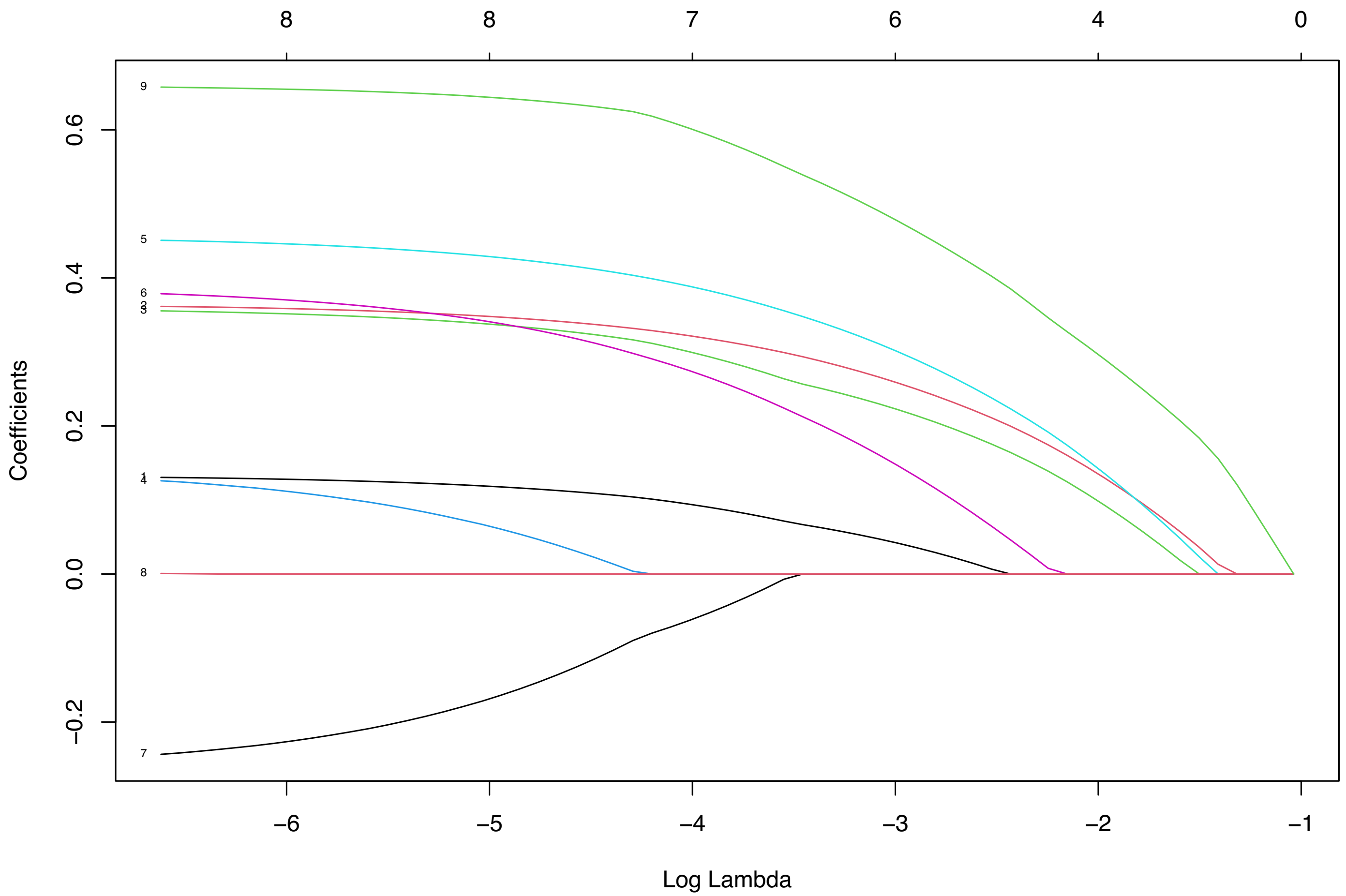
```
10 x 3 sparse Matrix of class "dgCMatrix"
                     lasso       ridge     logistic
(Intercept)    -0.70977228 -0.71220689 -0.878545196
sbp             .           0.10221203  0.133308398
tobacco         0.18103811  0.21846208  0.364577926
ldl             0.14726886  0.18656817  0.360180594
adiposity       .           0.11163533  0.144616485
famhistPresent  0.22246385  0.23181050  0.456537713
typea           0.01954765  0.13189202  0.388725509
obesity         .          -0.03579032 -0.265082072
alcohol         .           0.01941844  0.002978424
age             0.43990121  0.28192570  0.660695163
```

Elastic net logistic regression

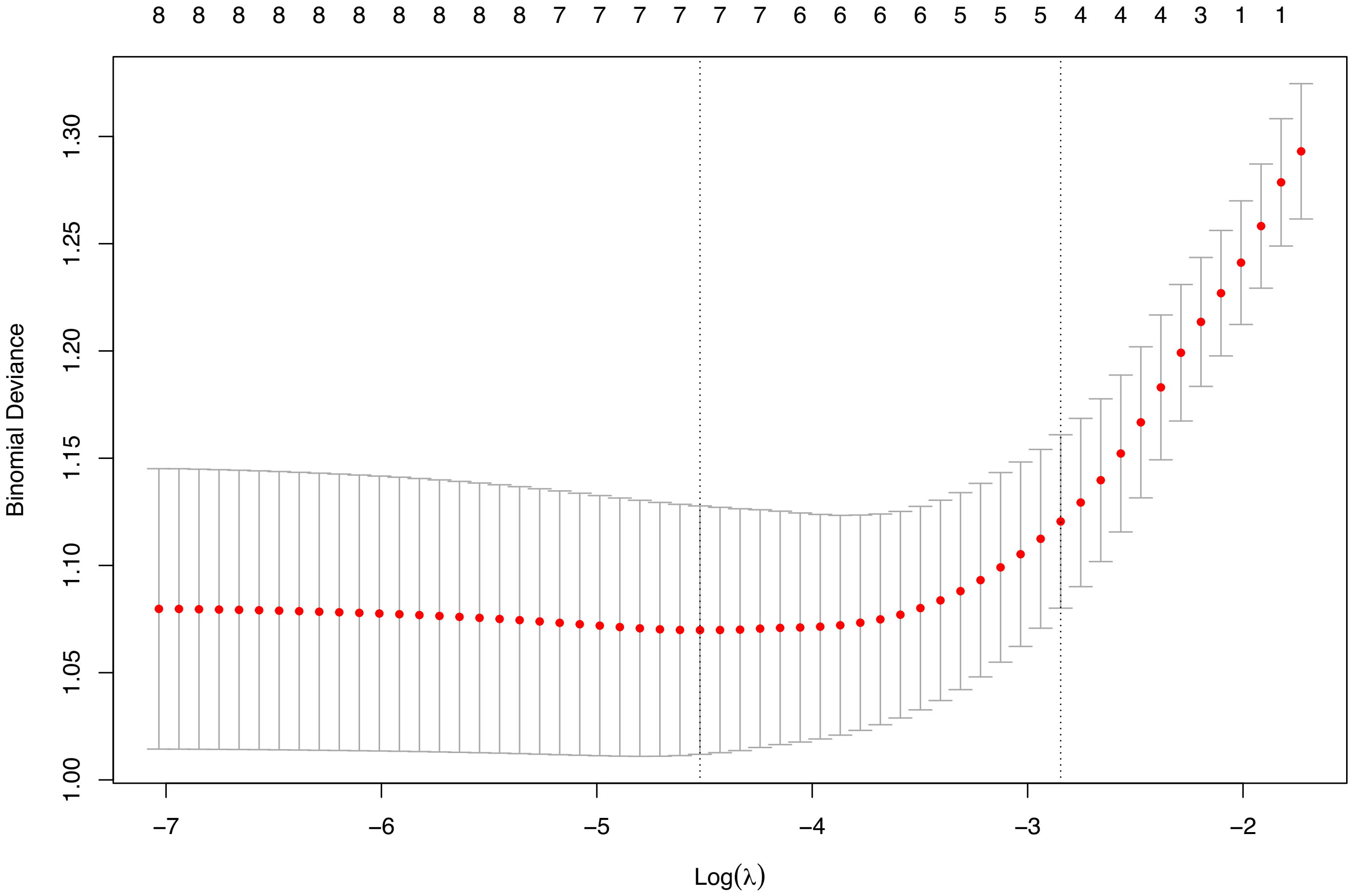
```r
cbind(1:9,colnames(xss))
```

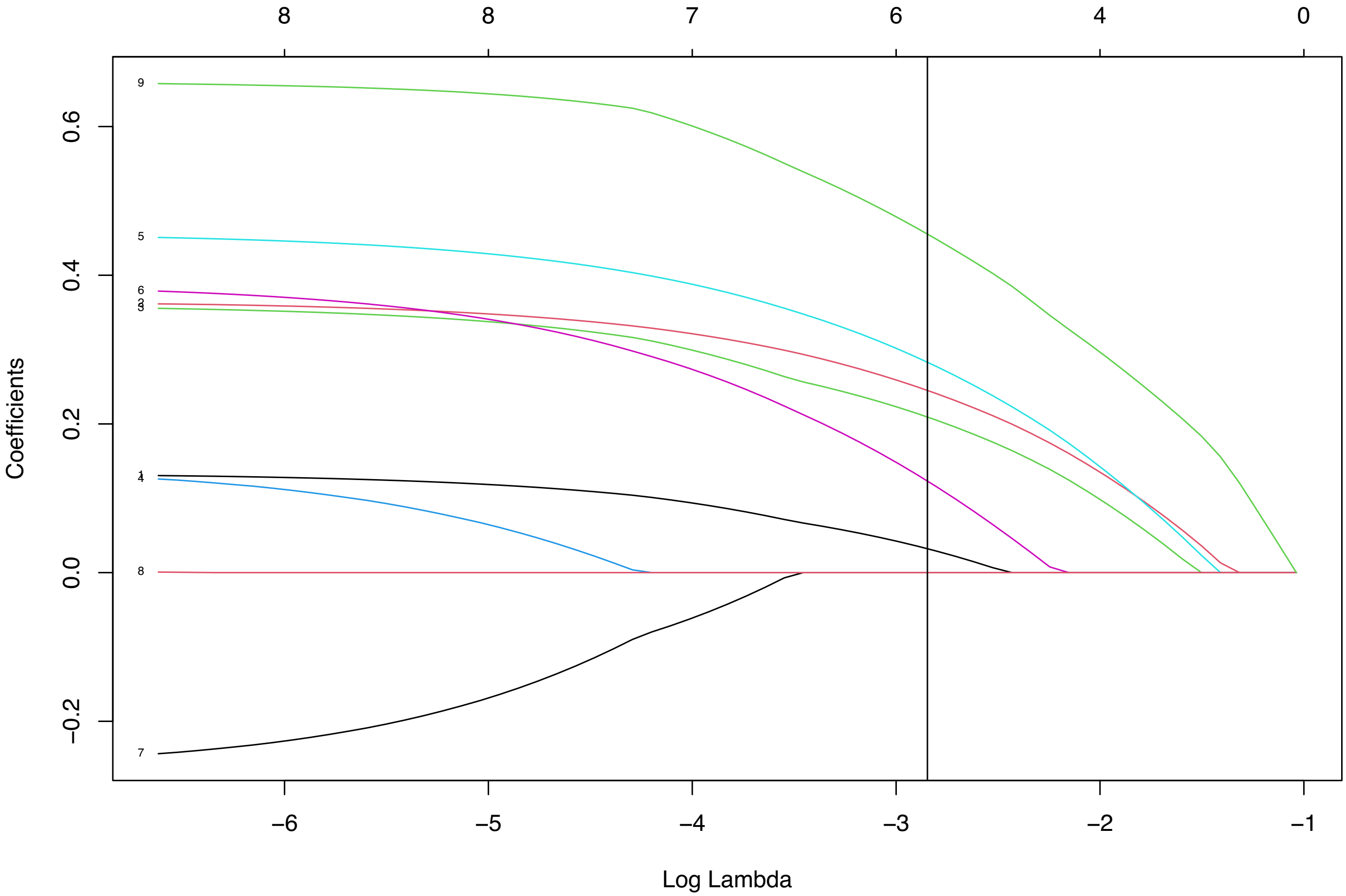
```
      [,1] [,2]
 [1,] "1"  "sbp"
 [2,] "2"  "tobacco"
 [3,] "3"  "ldl"
 [4,] "4"  "adiposity"
 [5,] "5"  "famhistPresent"
 [6,] "6"  "typea"
 [7,] "7"  "obesity"
 [8,] "8"  "alcohol"
 [9,] "9"  "age"
```

```r
elfit=glmnet(x=xss,y=ys,alpha=0.5,standardize=FALSE,family=
```


8
8
7
6
4
0
Coefficients
0.6
0.4
0.2
0.0
−0.2
9
5
6
2
3
1
4
8
7
−6
−5
−4
−3
−2
−1
Log Lambda

```
[1] "The lamda giving the smallest CV error 0.0108769601280
[1] "The 1sd err method lambda 0.0580470647530891"
```




8
8
7
6
4
0
0.6
0.4
0.2
0.0
−0.2
Coefficients
9
5
6
2
3
1
4
8
7
−6
−5
−4
−3
−2
−1
Log Lambda

```
10 x 4 sparse Matrix of class "dgCMatrix"
                   elastic       lasso       ridge       logi
(Intercept)    -0.73777844 -0.70977228 -0.71220689 -0.87854
sbp             0.03217102  .           0.10221203  0.13330
tobacco         0.24511842  0.18103811  0.21846208  0.36457
ldl             0.20932546  0.14726886  0.18656817  0.36018
adiposity       .           .           0.11163533  0.14461
famhistPresent  0.28303831  0.22246385  0.23181050  0.45653
typea           0.12327428  0.01954765  0.13189202  0.38872
obesity         .           .          -0.03579032 -0.26508
alcohol         .           .           0.01941844  0.00297
age             0.45547081  0.43990121  0.28192570  0.66069
```

corr

Computational details for the glmnet

Read for yourself!

(HTW 3.7)

`glmnet` is the implementation in R of the elastic net from HTW-book, and the package is maintained by Trevor Hastie.

The package fits generalized linear models using penalized maximum likelihood of elastic net type (lasso and ridge are special cases).

The logistic lasso is fitted using a quadratic approximation for the negative log-likelihood in a “proximal-Newton iterative approach”.

Software links

- R glmnet on CRAN with resources.
  - Getting started
  - GLM with glmnet

For Python there are different options.

- Python glmnet is recommended by Hastie et al.
- scikit-learn (seems to mostly be for regression? is there lasso for classification here?)

glmnet inputs

```r
glmnet(x, y,
 family = c("gaussian", "binomial", "poisson", "multinomial
 weights = NULL, offset = NULL, alpha = 1, nlambda = 100,
 lambda.min.ratio = ifelse(nobs < nvars, 0.01, 1e-04),
 lambda = NULL, standardize = TRUE, intercept = TRUE,
 thresh = 1e-07, dfmax = nvars + 1,
 pmax = min(dfmax * 2 + 20, nvars),
 exclude = NULL, penalty.factor = rep(1, nvars),
 lower.limits = -Inf, upper.limits = Inf, maxit = 1e+05,
 type.gaussian = ifelse(nvars < 500, "covariance", "naive")
 type.logistic = c("Newton", "modified.Newton"),
 standardize.response = FALSE,
 type.multinomial = c("ungrouped","grouped"),
 relax = FALSE, trace.it = 0, ...)
```

cv.glmnet inputs

```
cv.glmnet(x, y, weights = NULL, offset = NULL, lambda = NUL
  type.measure = c("default", "mse", "deviance", "class", "
  nfolds = 10, foldid = NULL,
  alignment = c("lambda", "fraction"), grouped = TRUE,
  keep = FALSE, parallel = FALSE,
  gamma = c(0, 0.25, 0.5, 0.75, 1), relax = FALSE, trace.it
```

type.measure defaults to deviance (accoring to help(cv.glmnet)).
The last is for Cox models.

Family

we have only covered `gaussian` (the default) and `binomial`. Each family has implemented the deviance measure. Poisson regression and Cox proportional hazard (survival analysis) is also implemented in glmnet.

Penalties

The elastic net is implemented, with three possible adjustment parameters.

$$\text{minimize}_{\beta_0,\beta}\{-\frac{1}{N}l(y;\beta_0,\beta)+\lambda\sum_{j=1}^{p}\gamma_j((1-\alpha)\beta_j^2+\alpha|\beta_j|)\}$$

- λ: the penalty, default a grid of 100 values is chosen, to cover the lasso path on the log scale.
- α: elastic net parameter $\in [0,1]$. This is usually manually selected by a grid search over 3-5 values. Default is $\alpha = 1$ (lasso), and with $\alpha = 0$ we get ridge.
- γ_j: penalty modifier for each covariate to be able to always include ($\gamma_j == 0$), or exclude ($\gamma_j =$ Inf), or give individual penalty modifications. Default $\lambda_j = 1$.

For the λ penalty the maximal value is for

- linear regression: $\lambda_{max} = \text{max}_j|\hat{\beta}_{LS,j}|$ (standardized coefficients) or, should there also be a factor 1/N?
- logistic regression: $\lambda_{max} = \text{max}_j|x_j^T(y - \bar{p})|$ where $\bar{p}$ is the mean case rate.

Additional modifications

- Coefficient bounds can be set (possible since coordinate descent is used)
- Some coefficients can be excluded from the penalization (than thus forced in).
- Offset can be added (popular if rate models for Poisson is used)
- For binary and multinomial data factors or matrices can be input.
- Sparse matrices with covariates can be supplied.

Lasso variants

Elastic net is already in glmnet (alpha-parameter).

Other lasso variants have their own R packages:

- The group lasso https://cran.r-project.org/web/packages/grplasso/grplasso.pdf
- The fused lasso https://cran.r-project.org/web/packages/genlasso/genlasso.pdf
- The sparse group lasso https://arxiv.org/pdf/2208.02942 and https://cran.r-project.org/web/packages/sparsegl/vignettes/sparsegl.html
- Bayesian lasso blasso function for normal data in package monomvn https://rdrr.io/cran/monomvn/man/monomvn-package.html
- Elastic net for ordinal data: https://cran.r-project.org/web/packages/ordinalNet/ordinalNet.pdf

Use mattelab to help eachother with solution for python?

Exercises

This week the best way to spend the time is to work on the Data Analysis Project 1.

But, also good to study the R-code for the South African heart disease example, and make some changes.

Smart: save this file as an .Rmd file and then run `purl(file.Rmd)` to produce a file with only the R-commands. (At the html-version you choose Code-Download Rmd on the top of the file).

- Change the CV criterion to auc and to class. Are there changes to what is the best choice for λ?

If you want to prepare
↓ for w6!

Supplemental sources useful for week 6 (see also the section on "Preparing for inference for the lasso and ridge")

- Bootstrap confidence intervals in the master thesis of Lene Tillerli Omdal Section 3.6.2 and teaching material from TMA4300 - see the wikipage for that course.
- Short note on multiple hypothesis testing in TMA4267 Linear Statistical Models, Kari K. Halle, Øyvind Bakke and Mette Langaas, March 15, 2017.