QCB 508 – Week 11

John D. Storey

Spring 2017

Contents

2

High-Dimensional Inference **2**
- Definition . . . 2
- Examples . . . 3
- HD Gene Expression Data . . . 4
- Many Responses Model . . . 5
- HD SNP Data . . . 7
- Many Regressors Model . . . 8
- Goals . . . 8
- Challenges . . . 9

Many Responses Model **9**

Shrinkage and Empirical Bayes **9**
- Estimating Several Means . . . 9
- Usual MLE . . . 9
- Loss Function . . . 9
- Stein's Paradox . . . 10
- Inverse Regression Approach . . . 14
- Empirical Bayes Estimate . . . 17
- EB for a Many Responses Model . . . 18

Multiple Testing **19**
- Motivating Example . . . 19
- Challenges . . . 20
- Outcomes . . . 21
- Error Rates . . . 21
- Bonferroni Correction . . . 22
- False Discovery Rate . . . 22
- Point Estimate . . . 23
- Adaptive Threshold . . . 24
- Conservative Properties . . . 24
- Q-Values . . . 24
- Bayesian Mixture Model . . . 25
- Bayesian-Frequentist Connection . . . 26
- Local FDR . . . 26

Many Regressors Model **27**

Ridge Regression **27**
Motivation . . . 27
Optimization Goal . . . 27
Solution . . . 28
Preprocessing . . . 28
Shrinkage . . . 28
Example . . . 28
Existence of Solution . . . 29
Effective Degrees of Freedom . . . 29
Bias and Covariance . . . 29
Ridge vs OLS . . . 30
Bayesian Interpretation . . . 30
Example: Diabetes Data . . . 30
GLMs . . . 32

Lasso Regression **33**
Motivation . . . 33
Optimization Goal . . . 33
Solution . . . 33
Preprocessing . . . 33
Bayesian Interpretation . . . 35
Inference . . . 35
GLMs . . . 36

Extras **36**
Source . . . 36
Session Information . . . 36

High-Dimensional Inference

Definition

High-dimesional inference is the scenario where we perform inference simultaneously on "many" paramaters.

"Many" can be as few as three parameters (which is where things start to get interesting), but in modern applications this is typically on the order of thousands to billions of paramaters.

High-dimesional data is a data set where the number of variables measured is many.

Large same size data is a data set where few variables are measured, but many observations are measured.

Big data is a data set where there are so many data points that it cannot be managed straightforwardly in memory, but must rather be stored and accessed elsewhere. Big data can be high-dimensional, large sample size, or both.

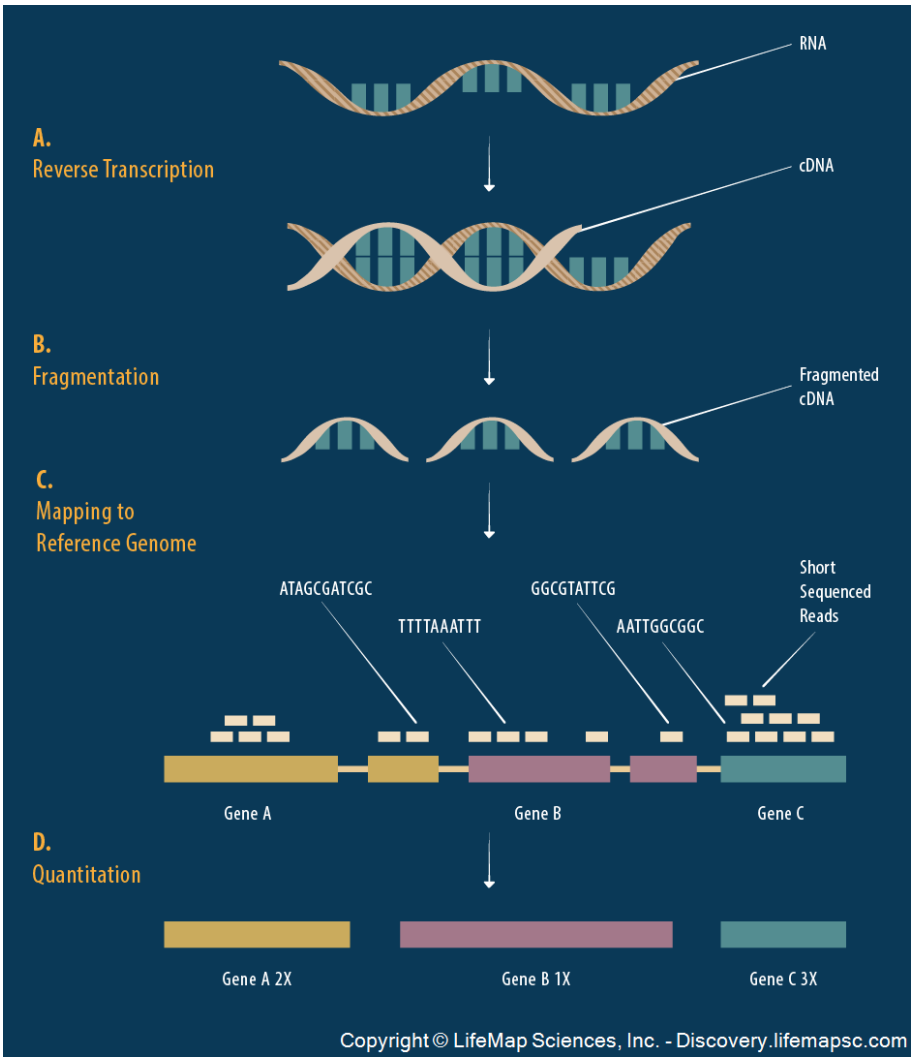
We will abbreviate high-dimensional with **HD**.

Examples

In all of these examples, many measurements are taken and the goal is often to perform inference on many paramaters simultaneously.

- Spatial epidemiology
- Environmental monitoring
- Internet user behavior
- Genomic profiling
- Neuroscience imaging
- Financial time series

HD Gene Expression Data



It's possible to measure the level of gene expression – how much mRNA is being transcribed – from thousands of cell simultaneously in a single biological sample.

Typically, gene expression is measured over varying biological conditions, and the goal is to perform inference on the relationship between expression and the varying conditions.

This results in thousands of simultaneous inferences.

The typical sizes of these data sets are 1000 to 50,000 genes and 10 to 1000 observations.

The gene expression values are typically modeled as approximately Normal or overdispersed Poisson.

There is usually shared signal among the genes, and there are often unobserved latent variables.

$$\underset{\text{genes}}{} \quad \underset{\boldsymbol{Y}_{m\times n}}{\underset{\text{observations}}{\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{bmatrix}}}$$

$$\underset{\boldsymbol{X}_{d\times n} \text{ study design}}{\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{d1} & x_{d2} & \cdots & x_{dn} \end{bmatrix}}$$

The $\boldsymbol{Y}$ matrix contains gene expression measurements for m genes (rows) by n observations (columns). The values y_{ij} are either in $\mathbb{R}$ or $\mathbb{Z}^+ = \{0, 1, 2, \ldots\}$.

The $\boldsymbol{X}$ matrix contains the study design of d explanatory variables (rows) by the n observations (columns).

Note that $m \gg n \gg d$.

Many Responses Model

Gene expression is an example of what I call the **many responses model**.

We're interested in performing simultaneous inference on d paramaters for each of m models such as:

$$\begin{aligned} \boldsymbol{Y}_1 &= \boldsymbol{\beta}_1 \boldsymbol{X} + \boldsymbol{E}_1 \\ \boldsymbol{Y}_2 &= \boldsymbol{\beta}_2 \boldsymbol{X} + \boldsymbol{E}_2 \\ &\vdots \\ \boldsymbol{Y}_m &= \boldsymbol{\beta}_m \boldsymbol{X} + \boldsymbol{E}_m \end{aligned}$$

For example, $\boldsymbol{Y}_1 = \boldsymbol{\beta}_1 \boldsymbol{X} + \boldsymbol{E}_1$ is vector notation of (in terms of observations j):

$$\{Y_{1j} = \beta_{11}X_{1j} + \beta_{12}X_{2j} + \cdots + \beta_{1d}X_{dj} + E_{1j}\}_{j=1}^{n}$$

We have made two changes from last week:

1. We have transposed $\boldsymbol{X}$ and $\boldsymbol{\beta}$.
2. We have changed the number of explanatory variables from p to d.

Let $\boldsymbol{B}_{m\times d}$ be the matrix of parameters (β_{ik}) relating the m response variables to the d explanatory variables. The full HD model is

$$\begin{array}{ccccccc} \boldsymbol{Y}_{m\times n} & = & \boldsymbol{B}_{m\times d} & \boldsymbol{X}_{d\times n} & + & \boldsymbol{E}_{m\times n} \\ \begin{bmatrix} Y_{i1} & \cdots & Y_{in} \end{bmatrix} & = & \begin{bmatrix} \beta_{i1} & \beta_{id} \end{bmatrix} & \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ X_{d1} & \cdots & X_{dn} \end{bmatrix} & + & \begin{bmatrix} E_{i1} & \cdots & E_{in} \end{bmatrix} \end{array}$$

Note that if we make OLS assumptions, then we can calculate:

$$\hat{\boldsymbol{B}}^{\text{OLS}} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$$

$$\hat{\boldsymbol{Y}} = \hat{\boldsymbol{B}}\boldsymbol{X} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}$$

so here the projection matrix is $\boldsymbol{P} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{X}$ and acts from the RHS, $\hat{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{P}$.

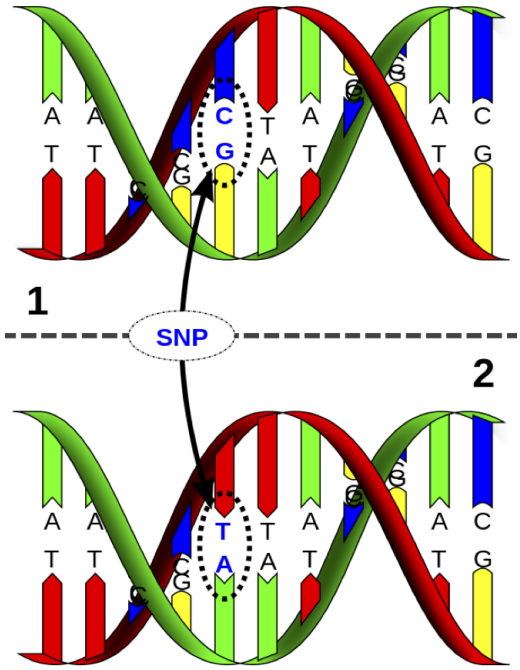
We will see this week and next that $\hat{\boldsymbol{B}}^{\text{OLS}}$ has nontrivial drawbacks. Thefore, we will be exploring other ways of estimating $\boldsymbol{B}$.

We of course aren't limited to OLS models. We could consider the many response GLM:

$$g\left(\mathrm{E}\left[\boldsymbol{Y}_{m\times n}|\,\boldsymbol{X}\right]\right) = \boldsymbol{B}_{m\times d}\boldsymbol{X}_{d\times n}$$

and we could even replace $\boldsymbol{B}_{m\times d}\boldsymbol{X}_{d\times n}$ with d smoothers for each of the m response variable.

HD SNP Data



It is possible to measure single nucleotide polymorphisms at millions of locations across the genome.

The base (A, C, G, or T) is measured from one of the strands.

For example, on the figure to the left, the individual is heterozygous CT at this SNP location.

$$\begin{array}{c} \boldsymbol{X}_{m\times n} \\ \text{individuals} \\ \text{SNPs} \end{array} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

$$\begin{array}{c} \boldsymbol{y}_{1\times n} \text{ trait} \\ \end{array} \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \end{bmatrix}$$

The $\boldsymbol{X}$ matrix contains SNP genotypes for m SNPs (rows) by n individuals

(columns). The values $x_{ij} \in \{0, 1, 2\}$ are conversions of genotypes (e.g., CC, CT, TT) to counts of one of the alleles.

The $\boldsymbol{y}$ vector contains the trait values of the n individuals.

Note that $m \gg n$.

Many Regressors Model

The SNP-trait model is an example of what I call the **many regressors model**. A single model is fit of a response variable on many regressors (i.e., explanatory variables) simultaneously.

This involves simultaneously inferring m paramaters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_m)$ in models such as:

$$\boldsymbol{Y} = \alpha \mathbf{1} + \boldsymbol{\beta} \boldsymbol{X} + \boldsymbol{E}$$

which is an n-vector with component j being:

$$Y_j = \alpha + \sum_{i=1}^{m} \beta_i X_{ij} + E_j$$

As with the many responses model, we do not need to limit the model to the OLS type where the response variable is approximately Normal distributed. Instead we can consider more general models such as

$$g\left(\mathrm{E}\left[\boldsymbol{Y} \middle| \boldsymbol{X}\right]\right) = \alpha \mathbf{1} + \boldsymbol{\beta} \boldsymbol{X}$$

for some link function $g(\cdot)$.

Goals

In both types of models we are interested in:

- Forming point estimates
- Testing statistical hypothesis
- Calculating posterior distributions
- Leveraging the HD data to increase our power and accuracy

Sometimes we are also interested in confidence intervals in high dimensions, but this is less common.

Challenges

Here are several of the new challenges we face when analyzing high-dimensional data:

- Standard estimation methods may be suboptimal in high dimensions
- New measures of significance are needed
- There may be dependence and latent variables among the high-dimensional variables
- The fact that $m \gg n$ poses challenges, especially in the many regressors model

HD data provide new challenges, but they also provide opportunities to model variation in the data in ways not possible for low-dimensional data.

Many Responses Model

Shrinkage and Empirical Bayes

Estimating Several Means

Let's start with the simplest *many responses model* where there is only an interncept and only one observation per variable. This means that $n = 1$ and $d = 1$ where $\boldsymbol{X} = 1$.

This model can be written as $Y_i \sim \text{Normal}(\beta_i, 1)$ for the $i = 1, 2, \ldots, m$ response variables. Suppose also that $Y_1, Y_2, \ldots, Y_m$ are jointly independent.

Let's assume that $\beta_1, \beta_2, \ldots, \beta_m$ are *fixed, nonrandom parameters.*

Usual MLE

The usual estimates of β_i are to set

$$\hat{\beta}_i^{\text{MLE}} = \boldsymbol{Y}_i.$$

This is also the OLS solution.

Loss Function

Suppose we are interested in the simultaneous loss function

$$L(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \sum_{i=1} (\beta_i - \hat{\beta}_i)^2$$

with risk $R(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \mathrm{E}[L(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})]$.

Stein's Paradox

Consider the following **James-Stein estimator**:

$$\hat{\beta}_i^{\mathrm{JS}} = \left(1 - \frac{m-2}{\sum_{k=1}^{m} Y_k^2}\right) Y_i.$$

In a shocking result called **Stein's paradox**, it was shown that when $m \geq 3$ then

$$R\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{\mathrm{JS}}\right) < R\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{\mathrm{MLE}}\right).$$

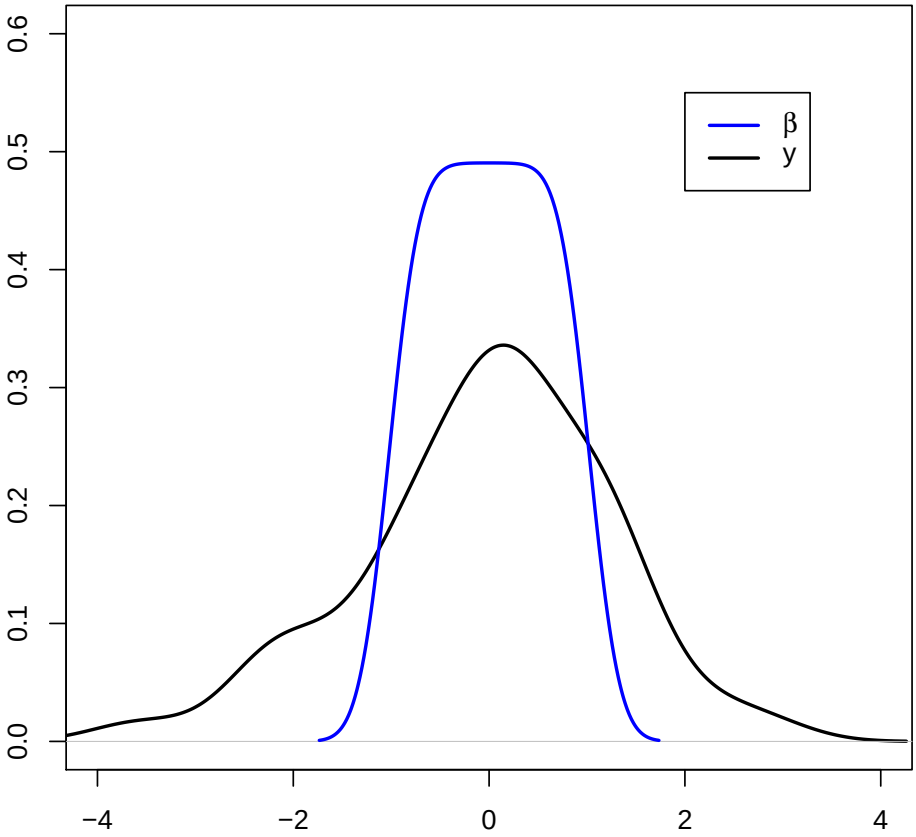
This means that the usual MLE is dominated by this JS estimator for any, even nonrandom, configuration of $\beta_1, \beta_2, \ldots, \beta_m$!

What is going on?

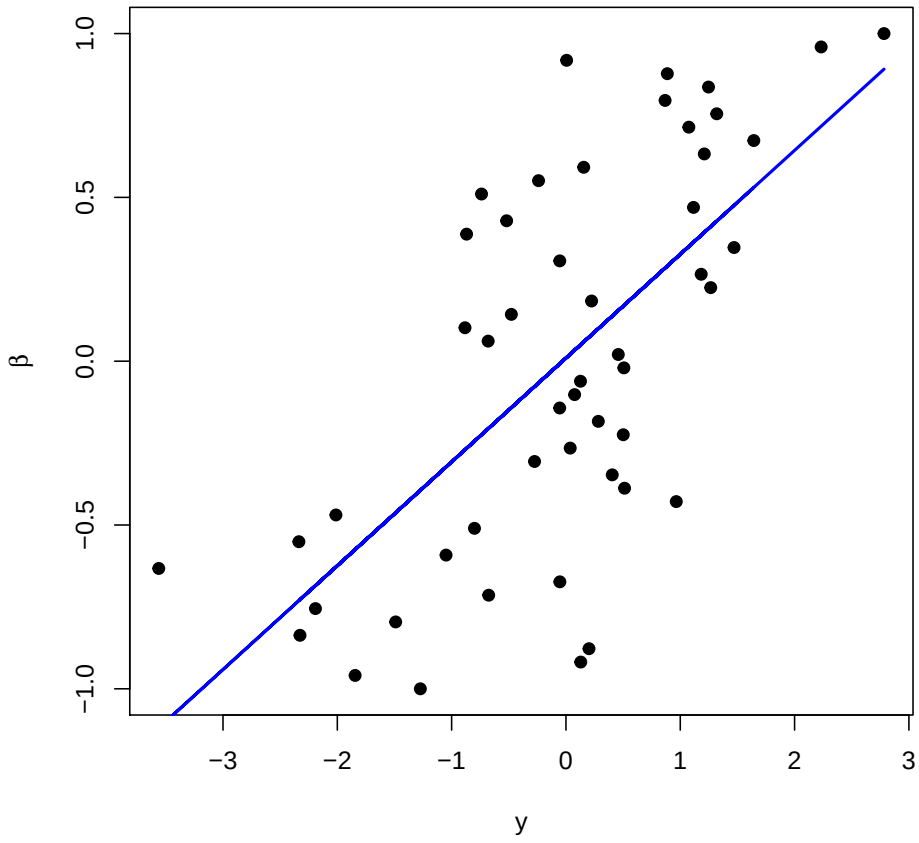
Let's first take a *linear regression* point of view to better understand this paradox.

Then we will return to the *empirical Bayes* example from earlier.

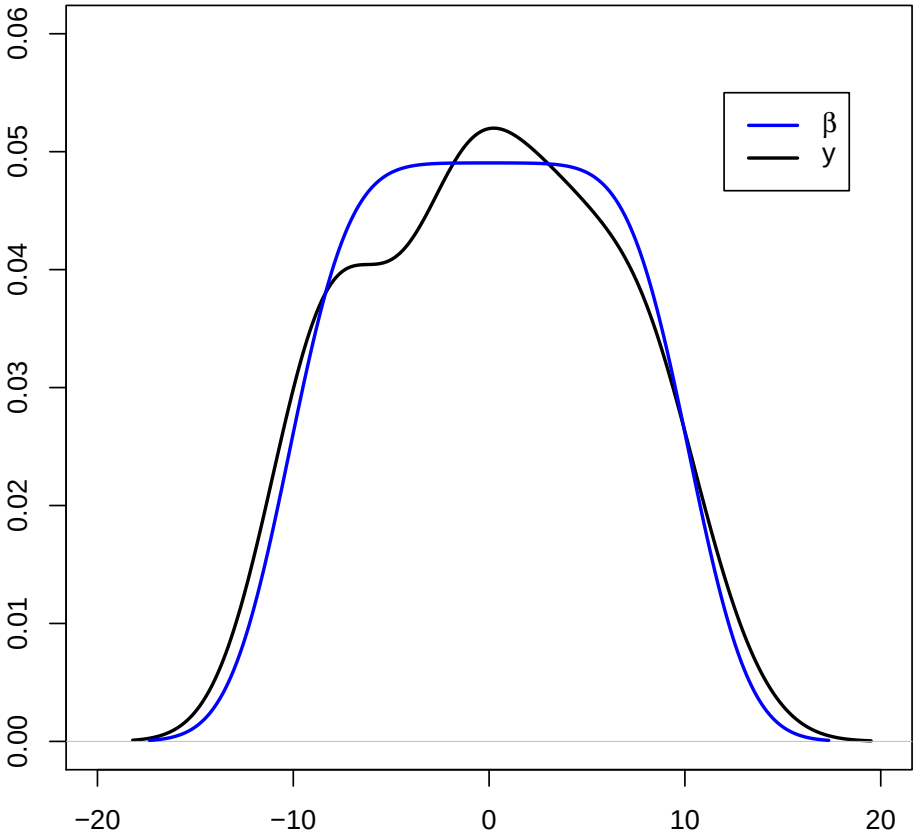
```
> beta <- seq(-1, 1, length.out=50)
> y <- beta + rnorm(length(beta))
```



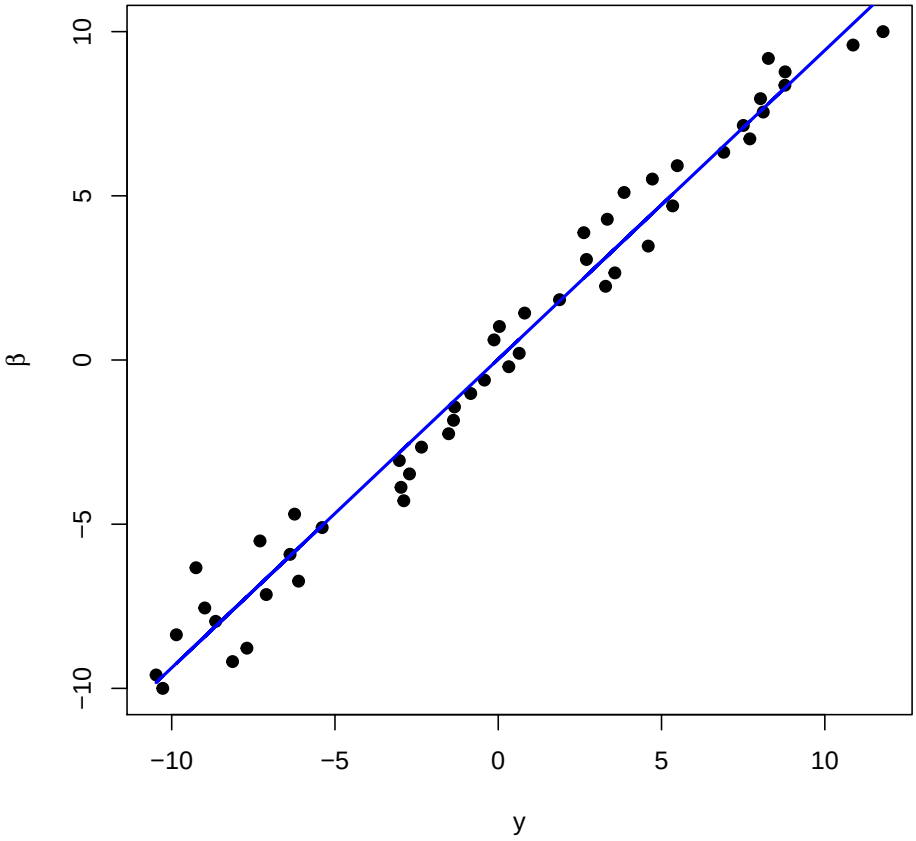
The blue line is the least squares regression line.




```
> beta <- seq(-10, 10, length.out=50)
> y <- beta + rnorm(length(beta))
```



The blue line is the least squares regression line.



Inverse Regression Approach

While $Y_i = \beta_i + E_i$ where $E_i \sim \text{Normal}(0, 1)$, it is also the case that $\beta_i = Y_i - E_i$ where $-E_i \sim \text{Normal}(0, 1)$.

Even though we're assuming the β_i are fixed, suppose we imagine for the moment that the β_i are random and take a least squares appraoch. We will try to estimate the linear model

$$\text{E}[\beta_i | Y_i] = a + bY_i.$$

Why would we do this? The loss function is

$$\sum_{i=1}^{m} (\beta_i - \hat{\beta}_i)^2$$

so it makes sense to estimate β_i by setting $\hat{\beta}_i$ to a regression line.

The least squares solution tells us to set

$$\begin{aligned}\hat{\beta}_i &= \hat{a} + \hat{b}Y_i \\ &= (\bar{\beta} - \hat{b}\bar{Y}) + \hat{b}Y_i \\ &= \bar{\beta} + \hat{b}(Y_i - \bar{Y})\end{aligned}$$

where

$$\hat{b} = \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})(\beta_i - \bar{\beta})}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}.$$

We can estimate $\bar{\beta}$ with $\bar{Y}$ since $\mathrm{E}[\bar{\beta}] = \mathrm{E}[\bar{Y}]$.
We also need to find an estimate of $\sum_{i=1}^{m}(Y_i - \bar{Y})(\beta_i - \bar{\beta})$. Note that

$$\beta_i - \bar{\beta} = Y_i - \bar{Y} - (E_i + \bar{E})$$

so that

$$\begin{aligned}\sum_{i=1}^{m}(Y_i - \bar{Y})(\beta_i - \bar{\beta}) = &\sum_{i=1}^{m}(Y_i - \bar{Y})(Y_i - \bar{Y}) \\ &+ \sum_{i=1}^{m}(Y_i - \bar{Y})(E_i - \bar{E})\end{aligned}$$

Since $Y_i = \beta_i + E_i$ it follows that

$$\mathrm{E}\left[\sum_{i=1}^{m}(Y_i - \bar{Y})(E_i - \bar{E})\right] = \mathrm{E}\left[\sum_{i=1}^{m}(E_i - \bar{E})(E_i - \bar{E})\right] = m - 1.$$

Therefore,

$$\mathrm{E}\left[\sum_{i=1}^{m}(Y_i - \bar{Y})(\beta_i - \bar{\beta})\right] = \mathrm{E}\left[\sum_{i=1}^{m}(Y_i - \bar{Y})^2 - (m-1)\right].$$

This yields

$$\hat{b} = \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2 - (m-1)}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2} = 1 - \frac{m-1}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}$$

and

$$\hat{\beta}_i^{\mathrm{IR}} = \bar{Y} + \left(1 - \frac{m-1}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}\right)(Y_i - \bar{Y})$$

If instead we had started with the no intercept model

$$\mathrm{E}[\beta_i | Y_i] = bY_i.$$

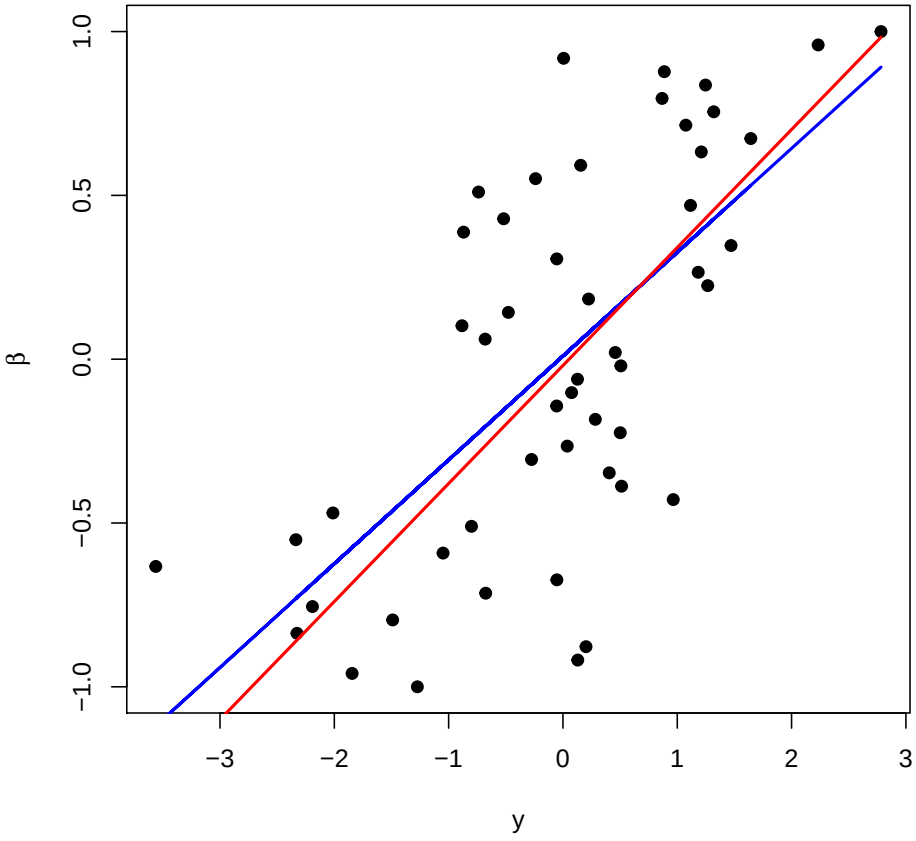
we would have ended up with

$$\hat{\beta}_i^{\mathrm{IR}} = \left(1 - \frac{m-1}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}\right) Y_i$$

In either case, it can be shown that

$$R\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{\mathrm{IR}}\right) < R\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{\mathrm{MLE}}\right).$$

The blue line is the least squares regression line of β_i on Y_i, and the red line is $\hat{\beta}_i^{\mathrm{IR}}$.



Empirical Bayes Estimate

Suppose that $Y_i|\beta_i \sim \text{Normal}(\beta_i, 1)$ where these rv's are jointly independent. Also suppose that $\beta_i \overset{\text{iid}}{\sim} \text{Normal}(a, b^2)$. Taking the empirical Bayes approach, we get:

$$f(y_i; a, b) = \int f(y_i|\beta_i) f(\beta_i; a, b) d\beta_i \sim \text{Normal}(a, 1 + b^2).$$

$$\implies \hat{a} = \overline{Y},\ 1 + \hat{b}^2 = \frac{\sum_{k=1}^{m} (Y_k - \overline{Y})^2}{n}$$

$$\mathrm{E}[\beta_i|Y_i] = \frac{1}{1+b^2}a + \frac{b^2}{1+b^2}Y_i \implies$$

$$\begin{aligned}\hat{\beta}_i^{\mathrm{EB}} = \hat{\mathrm{E}}[\beta_i|Y_i] &= \frac{1}{1+\hat{b}^2}\hat{a} + \frac{\hat{b}^2}{1+\hat{b}^2}Y_i \\ &= \frac{m}{\sum_{k=1}^m (Y_k - \overline{Y})^2}\overline{Y} + \left(1 - \frac{m}{\sum_{k=1}^m (Y_k - \overline{Y})^2}\right) Y_i\end{aligned}$$

As with $\hat{\boldsymbol{\beta}}^{\mathrm{JS}}$ and $\hat{\boldsymbol{\beta}}^{\mathrm{IR}}$, we have

$$R\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{\mathrm{EB}}\right) < R\left(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}^{\mathrm{MLE}}\right).$$

EB for a Many Responses Model

Consider the *many responses model* where $\boldsymbol{Y}_i|\boldsymbol{X} \sim \mathrm{MVN}_n(\boldsymbol{\beta}_i\boldsymbol{X}, \sigma^2\boldsymbol{I})$ where the vectors $\boldsymbol{Y}_i|\boldsymbol{X}$ are jointly independent ($i = 1, 2, \ldots, m$). Here we've made the simplifying assumption that the variance σ^2 is equal across all responses, but this would not be generally true.

The OLS (and MLE) solution is

$$\hat{\boldsymbol{B}} = \boldsymbol{Y}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}.$$

Suppose we extend this so that $\boldsymbol{Y}_i|\boldsymbol{X}, \boldsymbol{\beta}_i \sim \mathrm{MVN}_n(\boldsymbol{\beta}_i\boldsymbol{X}, \sigma^2\boldsymbol{I})$ and $\boldsymbol{\beta}_i \overset{\text{iid}}{\sim} \mathrm{MVN}_d(\boldsymbol{u}, \boldsymbol{V})$.

Since $\hat{\boldsymbol{\beta}}_i|\boldsymbol{\beta}_i \sim \mathrm{MVN}_d(\boldsymbol{\beta}_i, \sigma^2(\boldsymbol{X}\boldsymbol{X}^T)^{-1})$, it follows that marginally

$$\hat{\boldsymbol{\beta}}_i \overset{\text{iid}}{\sim} \mathrm{MVN}_d(\boldsymbol{u}, \sigma^2(\boldsymbol{X}\boldsymbol{X}^T)^{-1} + \boldsymbol{V}).$$

Therefore,

$$\hat{\boldsymbol{u}} = \frac{\sum_{i=1}^m \hat{\boldsymbol{\beta}}_i}{m}$$

$$\hat{\boldsymbol{V}} = \hat{\mathrm{Cov}}\left(\hat{\boldsymbol{\beta}}\right) - \hat{\sigma}^2(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$$

where $\hat{\mathrm{Cov}}\left(\hat{\boldsymbol{\beta}}\right)$ is the $d \times d$ sample covariance (or MLE covariance) of the $\hat{\boldsymbol{\beta}}_i$ estimates.

Also, $\hat{\sigma}^2$ is obtained by averaging the estimate over all m regressions.

We then do inference based on the prior distribution $\boldsymbol{\beta}_i \overset{\text{iid}}{\sim} \text{MVN}_d(\hat{\boldsymbol{u}}, \hat{\boldsymbol{V}})$. The posterior distribution of $\boldsymbol{\beta}_i|\boldsymbol{Y}, \boldsymbol{X}$ is MVN with mean

$$\left(\frac{1}{\hat{\sigma}^2}(\boldsymbol{X}\boldsymbol{X}^T) + \hat{\boldsymbol{V}}^{-1}\right)^{-1}\left(\frac{1}{\hat{\sigma}^2}(\boldsymbol{X}\boldsymbol{X}^T)\hat{\boldsymbol{\beta}}_i + \hat{\boldsymbol{V}}^{-1}\hat{\boldsymbol{u}}\right)$$

and covariance

$$\left(\frac{1}{\hat{\sigma}^2}(\boldsymbol{X}\boldsymbol{X}^T) + \hat{\boldsymbol{V}}^{-1}\right)^{-1}.$$

Multiple Testing

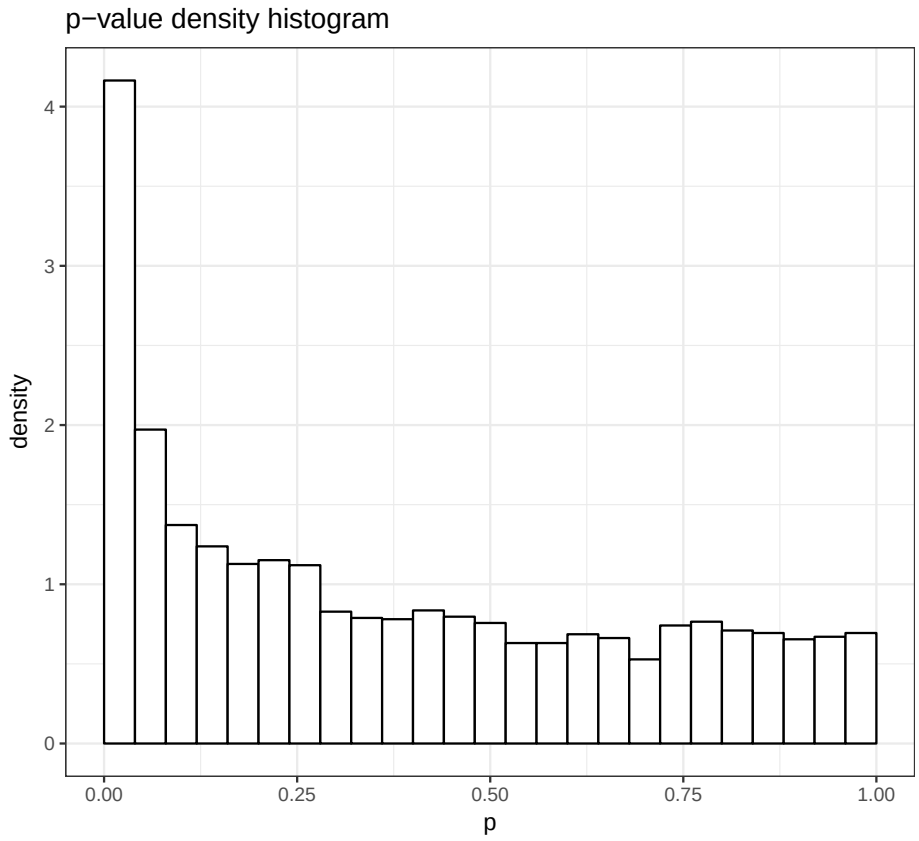
Motivating Example

Hedenfalk et al. (2001) *NEJM* measured gene expression in three different breast cancer tumor types. In your homework, you have analyzed these data and have specifically compared BRCA1 mutation positive tumors to BRCA2 mutation positive tumors.

The `qvalue` package has the p-values when testing for a difference in population means between these two groups (called "differential expression"). There are 3170 genes tested, resulting in 3170 p-values.

Note that this analysis is a version of the many responses model.

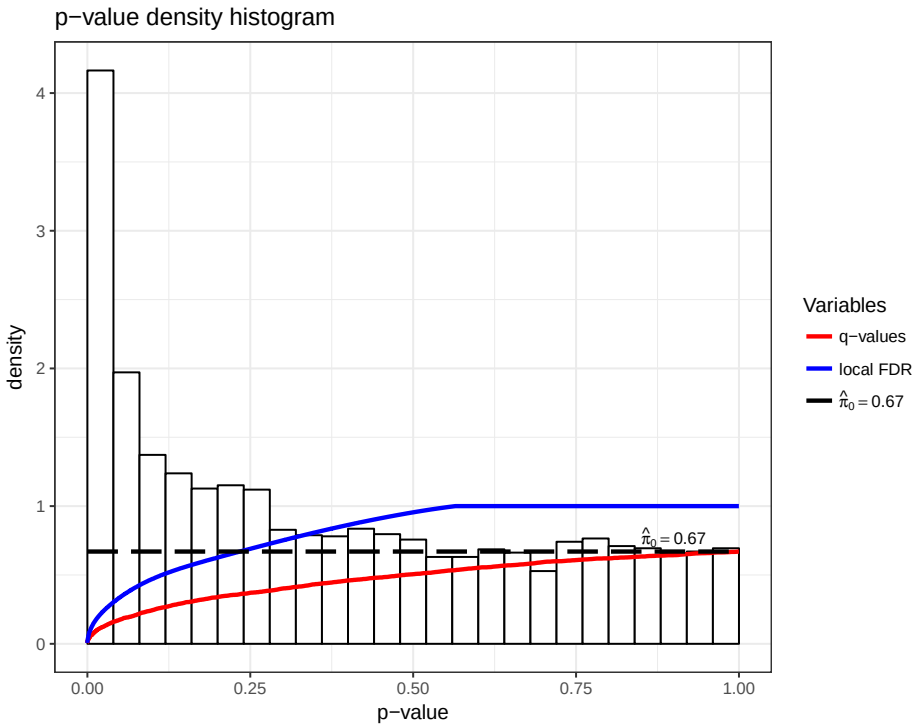
```
> library(qvalue)
> data(hedenfalk); df <- data.frame(p=hedenfalk$p)
> ggplot(df, aes(x = p)) +
+          ggtitle("p-value density histogram") +
+          geom_histogram(aes_string(y = '..density..'), colour = "black",
+                         fill = "white", binwidth = 0.04, center=0.02)
```



Challenges

- Traditional p-value thresholds such as 0.01 or 0.05 may result in too many false positives. For example, in the above example, a 0.05 threshold could result in 158 false positives.
- A careful balance of true positives and false positives must be achieved in a manner that is scientifically interpretable.
- There is information in the joint distribution of the p-values that can be leveraged.
- Dependent p-values may make this type of analysis especially difficult (next week's topic).

```
> qobj <- qvalue(hedenfalk$p)
> hist(qobj)
```



Outcomes

Possible outcomes from m hypothesis tests based on applying a significance threshold $0 < t \leq 1$ to their corresponding p-values.

	Not Significant	Significant	Total
Null True	U	V	m_0
Alternative True	T	S	m_1
Total	W	R	m

Error Rates

Suppose we are testing m hypotheses based on p-values $p_1, p_2, \ldots, p_m$.

Multiple hypothesis testing is the process of deciding which of these p-values should be called statistically significant.

This requires formulating and estimating a compound **error rate** that quantifies the quality of the decision.

Bonferroni Correction

The **family-wise error rate** is the probability of *any* false positive occurring among all tests called significant. The **Bonferroni correction** is a result that shows that utilizing a p-value threshold of α/m results in FWER $\leq \alpha$. Specifically,

$$\begin{aligned}\text{FWER} &\leq \Pr(\cup\{P_i \leq \alpha/m\}) \\ &\leq \sum_{i=1}^{m} \Pr(P_i \leq \alpha/m) = \sum_{i=1}^{m} \alpha/m = \alpha\end{aligned}$$

where the above probability calculations are done under the assumption that all H_0 are true.

False Discovery Rate

The **false discovery rate** (FDR) measures the proportion of Type I errors — or "false discoveries" — among all hypothesis tests called statistically significant. It is defined as

$$\text{FDR} = \text{E}\left[\frac{V}{R \vee 1}\right] = \text{E}\left[\frac{V}{R}\middle| R > 0\right] \Pr(R > 0).$$

This is less conservative than the FWER and it offers a clearer balance between true positives and false positives.

There are two other false discovery rate definitions, where the main difference is in how the $R = 0$ event is handled. These quantities are called the **positive false discovery rate** (pFDR) and the **marginal false discovery rate** (mFDR), defined as follows:

$$\text{pFDR} = \text{E}\left[\frac{V}{R}\middle| R > 0\right],$$

$$\text{mFDR} = \frac{\text{E}[V]}{\text{E}[R]}.$$

Note that pFDR = mFDR = 1 whenever all null hypotheses are true, whereas FDR can always be made arbitrarily small because of the extra term $\Pr(R > 0)$.

Point Estimate

Let FDR(t) denote the FDR when calling null hypotheses significant whenever $p_i \leq t$, for $i = 1, 2, \ldots, m$. For $0 < t \leq 1$, we define the following random variables:

$$V(t) = \#\{\text{true null } p_i : p_i \leq t\}$$
$$R(t) = \#\{p_i : p_i \leq t\}$$

In terms of these, we have

$$\text{FDR}(t) = \text{E}\left[\frac{V(t)}{R(t) \vee 1}\right].$$

For fixed t, the following defines a family of conservatively biased point estimates of FDR(t):

$$\hat{\text{FDR}}(t) = \frac{\hat{m}_0(\lambda) \cdot t}{[R(t) \vee 1]}.$$

The term $\hat{m}_0(\lambda)$ is an estimate of m_0, the number of true null hypotheses. This estimate depends on the tuning parameter λ, and it is defined as

$$\hat{m}_0(\lambda) = \frac{m - R(\lambda)}{(1 - \lambda)}.$$

Sometimes instead of m_0, the quantity

$$\pi_0 = \frac{m_0}{m}$$

is estimated, where simply

$$\hat{\pi}_0(\lambda) = \frac{\hat{m}_0(\lambda)}{m} = \frac{m - R(\lambda)}{m(1 - \lambda)}.$$

It can be shown that $\text{E}[\hat{m}_0(\lambda)] \geq m_0$ when the p-values corresponding to the true null hypotheses are Uniform(0,1) distributed (or stochastically greater).

There is an inherent bias/variance trade-off in the choice of λ. In most cases, when λ gets smaller, the bias of $\hat{m}_0(\lambda)$ gets larger, but the variance gets smaller.

Therefore, λ can be chosen to try to balance this trade-off.

Adaptive Threshold

If we desire a FDR level of α, it is tempting to use the p-value threshold

$$t^*_\alpha = \max\left\{t : \hat{\text{FDR}}(t) \leq \alpha\right\}$$

which identifies the largest estimated FDR less than or equal to α.

Conservative Properties

When the p-value corresponding to true null hypothesis are distributed iid Uniform(0,1), then we have the following two conservative properties.

$$\text{E}\left[\hat{\text{FDR}}(t)\right] \geq \text{FDR}(t)$$
$$\text{E}\left[\hat{\text{FDR}}(t^*_\alpha)\right] \leq \alpha$$

Q-Values

In single hypothesis testing, it is common to report the p-value as a measure of significance. The **q-value** is the FDR based measure of significance that can be calculated simultaneously for multiple hypothesis tests.

The p-value is constructed so that a threshold of α results in a Type I error rate $\leq \alpha$. Likewise, the q-value is constructed so that a threshold of α results in a FDR $\leq \alpha$.

Initially it seems that the q-value should capture the FDR incurred when the significance threshold is set at the p-value itself, $\text{FDR}(p_i)$. However, unlike Type I error rates, the FDR is not necessarily strictly increasing with an increasing significance threshold.

To accommodate this property, the q-value is defined to be the minimum FDR (or pFDR) at which the test is called significant:

$$\text{q-value}(p_i) = \min_{t \geq p_i} \text{FDR}(t)$$

or

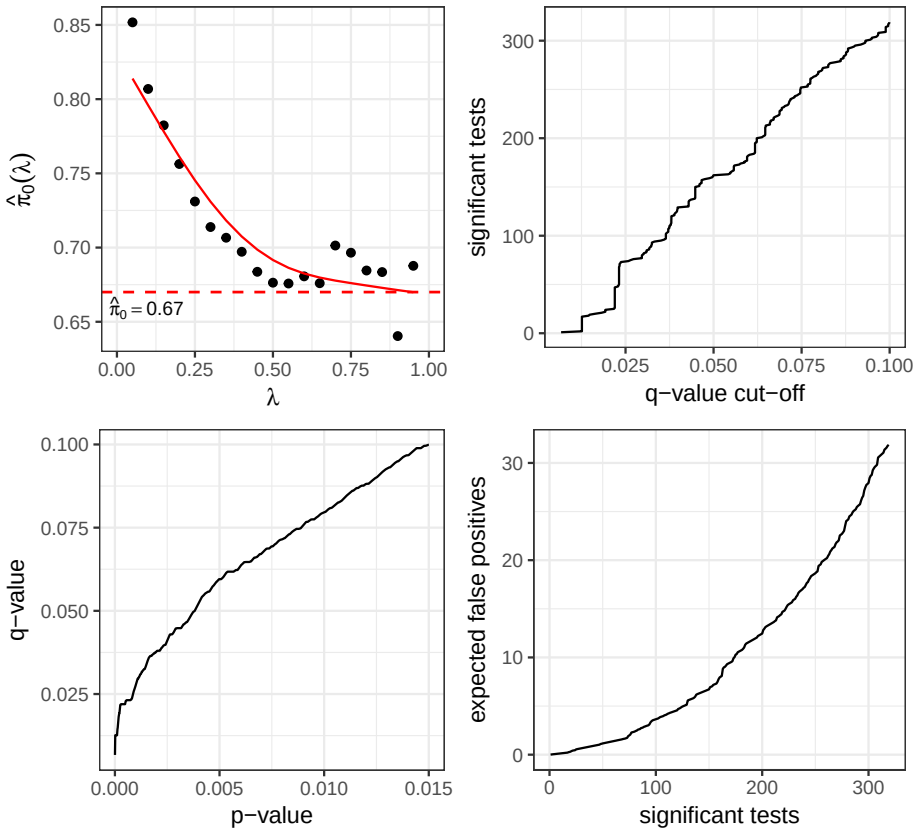
$$\text{q-value}(p_i) = \min_{t \geq p_i} \text{pFDR}(t).$$

To estimate this in practice, a simple plug-in estimate is formed, for example:

$$\hat{\text{q}}\text{-value}(p_i) = \min_{t \geq p_i} \hat{\text{FDR}}(t).$$

Various theoretical properties have been shown for these estimates under certain conditions, notably that the estimated q-values of the entire set of tests are simultaneously conservative as the number of hypothesis tests grows large.

```
> plot(qobj)
```



Bayesian Mixture Model

Let's return to the Bayesian classification set up from earlier. Suppose that

- $H_i = 0$ or 1 according to whether the ith null hypothesis is true or not
- $H_i \overset{\text{iid}}{\sim} \text{Bernoulli}(1-\pi_0)$ so that $\Pr(H_i = 0) = \pi_0$ and $\Pr(H_i = 1) = 1-\pi_0$
- $P_i | H_i \overset{\text{iid}}{\sim} (1 - H_i) \cdot F_0 + H_i \cdot F_1$, where F_0 is the null distribution and F_1 is the alternative distribution

Bayesian-Frequentist Connection

Under these assumptions, it has been shown that

$$\begin{aligned}\text{pFDR}(t) &= \text{E}\left[\frac{V(t)}{R(t)}\middle| R(t) > 0\right] \\ &= \Pr(H_i = 0 | P_i \leq t)\end{aligned}$$

where $\Pr(H_i = 0|P_i \leq t)$ is the same for each i because of the iid assumptions.

Under these modeling assumptions, it follows that

$$\text{q-value}(p_i) = \min_{t \geq p_i} \Pr(H_i = 0 | P_i \leq t)$$

which is a Bayesian analogue of the p-value — or rather a "Bayesian posterior Type I error rate".

Local FDR

In this scenario, it also follows that

$$\text{pFDR}(t) = \int \Pr(H_i = 0 | P_i = p_i) dF(p_i | p_i \leq t)$$

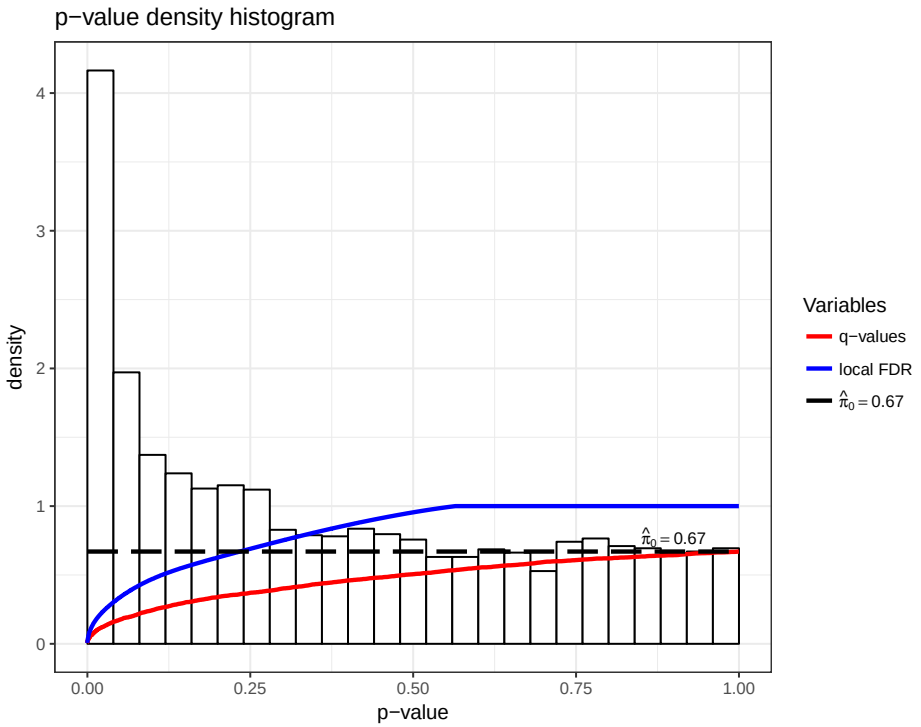
where $F = \pi_0 F_0 + (1 - \pi_0) F_1$.

This connects the pFDR to the **posterior error probability**

$$\Pr(H_i = 0 | P_i = p_i)$$

making this latter quantity sometimes interpreted as a **local false discovery rate**.

```
> hist(qobj)
```



Many Regressors Model

Ridge Regression

Motivation

Ridge regression is a technique for shrinking the coefficients towards zero in linear models.

It also deals with collinearity among explanatory variables. **Collinearity** is the presence of strong correlation among two or more explanatory variables.

Optimization Goal

Under the OLS model assumptions, ridge regression fits model by minimizing the following:

$$\sum_{j=1}^{n}\left(y_j - \sum_{i=1}^{m}\beta_i x_{ij}\right)^2 + \lambda\sum_{k=1}^{m}\beta_i^2.$$

Recall the ℓ_2 norm: $\sum_{k=1}^{m} \beta_i^2 = \|\boldsymbol{\beta}\|_2^2$. Sometimes ridge regression is called ℓ_2 penalized regression.

As with natural cubic splines, the paramater λ is a tuning paramater that controls how much shrinkage occurs.

Solution

The ridge regression solution is

$$\hat{\boldsymbol{\beta}}^{\text{Ridge}} = \boldsymbol{y}\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}\right)^{-1}.$$

As $\lambda \to 0$, the $\hat{\boldsymbol{\beta}}^{\text{Ridge}} \to \hat{\boldsymbol{\beta}}^{\text{OLS}}$.

As $\lambda \to \infty$, the $\hat{\boldsymbol{\beta}}^{\text{Ridge}} \to \mathbf{0}$.

Preprocessing

Implicitly...

We mean center $\boldsymbol{y}$.

We also mean center and standard deviation scale each explanatory variable. Why?

Shrinkage

When $\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{I}$, then

$$\hat{\beta}_j^{\text{Ridge}} = \frac{\hat{\beta}_j^{\text{OLS}}}{1+\lambda}.$$

This shows how ridge regression acts as a technique for shrinking regression coefficients towards zero. It also shows that when $\hat{\beta}_j^{\text{OLS}} \neq 0$, then for all finite λ, $\hat{\beta}_j^{\text{Ridge}} \neq 0$.

Example

```
> set.seed(508)
> x1 <- rnorm(20)
> x2 <- x1 + rnorm(20, sd=0.1)
> y <- 1 + x1 + x2 + rnorm(20)
> tidy(lm(y~x1+x2))
         term  estimate std.error statistic      p.value
1 (Intercept) 0.9647132 0.2036815 4.7363815 0.0001908984
2          x1 0.4927857 2.8104958 0.1753376 0.8628858043
3          x2 1.2599509 2.8876230 0.4363280 0.6680895121
> lm.ridge(y~x1+x2, lambda=1) # from MASS package
                  x1         x2
0.9486116 0.8252948 0.8751979
```

Existence of Solution

When $d > n$ or when there is high collinearity, then $(\boldsymbol{X}\boldsymbol{X}^T)^{-1}$ will not exist.

However, for $\lambda > 0$, it is always the case that $\left(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}\right)^{-1}$ exists.

Therefore, one can always compute a unique $\hat{\boldsymbol{\beta}}^{\text{Ridge}}$ for each $\lambda > 0$.

Effective Degrees of Freedom

Similarly to natural cubic splines, we can calculate an effective degrees of freedom by noting that:

$$\hat{\boldsymbol{y}} = \boldsymbol{y}\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}$$

The effective degrees of freedom is then the trace of the linear operator:

$$\text{tr}\left(\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}\right)$$

Bias and Covariance

Under the OLS model assumptions,

$$\text{Cov}\left(\hat{\boldsymbol{\beta}}^{\text{Ridge}}\right) = \sigma^2 \left(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}\right)^{-1} \boldsymbol{X}\boldsymbol{X}^T \left(\boldsymbol{X}\boldsymbol{X}^T + \lambda \boldsymbol{I}\right)^{-1}$$

and

$$\text{bias} = \mathrm{E}\left[\hat{\boldsymbol{\beta}}^{\text{Ridge}}\right] - \boldsymbol{\beta} = -\lambda\boldsymbol{\beta}\left(\boldsymbol{X}\boldsymbol{X}^T + \lambda\boldsymbol{I}\right)^{-1}.$$

Ridge vs OLS

When the OLS model is true, there exists a $\lambda > 0$ such that the MSE of the ridge estimate is lower than than of the OLS estimate:

$$\mathrm{E}\left[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{Ridge}}\|_2^2\right] < \mathrm{E}\left[\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{\text{OLS}}\|_2^2\right].$$

This says that by sacrificing some bias in the ridge estimator, we can obtain a smaller overall MSE, which is bias^2 + variance.

Bayesian Interpretation

The ridge regression solution is equivalent to maximizing

$$-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j - \sum_{i=1}^{m}\beta_i x_{ij}\right)^2 - \frac{\lambda}{2\sigma^2}\sum_{k=1}^{m}\beta_i^2$$

which means it can be interpreted as the MAP solution with a Normal prior on the β_i values.

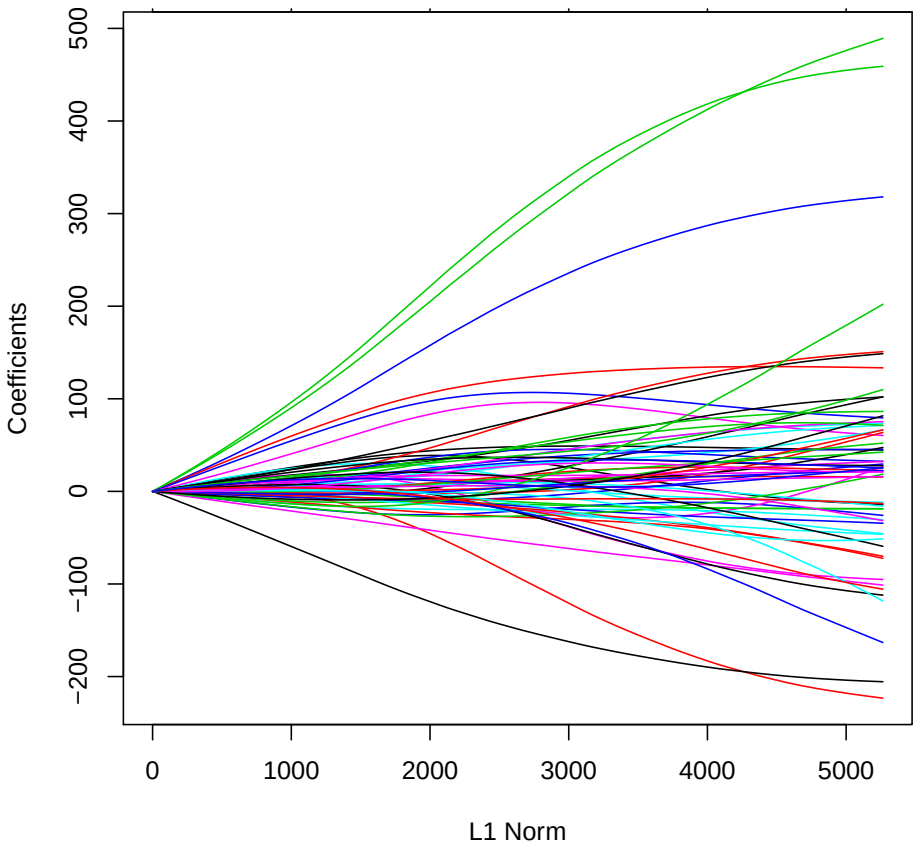
Example: Diabetes Data

```
> library(lars)
> data(diabetes)
> x <- diabetes$x2 %>% unclass() %>% as.data.frame()
> y <- diabetes$y
> dim(x)
[1] 442  64
> length(y)
[1] 442
> df <- cbind(x,y)
> names(df)
 [1] "age"     "sex"     "bmi"     "map"     "tc"
 [6] "ldl"     "hdl"     "tch"     "ltg"     "glu"
[11] "age^2"   "bmi^2"   "map^2"   "tc^2"    "ldl^2"
[16] "hdl^2"   "tch^2"   "ltg^2"   "glu^2"   "age:sex"
[21] "age:bmi" "age:map" "age:tc"  "age:ldl" "age:hdl"
[26] "age:tch" "age:ltg" "age:glu" "sex:bmi" "sex:map"
```

```
[31] "sex:tc"  "sex:ldl" "sex:hdl" "sex:tch" "sex:ltg"
[36] "sex:glu" "bmi:map" "bmi:tc"  "bmi:ldl" "bmi:hdl"
[41] "bmi:tch" "bmi:ltg" "bmi:glu" "map:tc"  "map:ldl"
[46] "map:hdl" "map:tch" "map:ltg" "map:glu" "tc:ldl"
[51] "tc:hdl"  "tc:tch"  "tc:ltg"  "tc:glu"  "ldl:hdl"
[56] "ldl:tch" "ldl:ltg" "ldl:glu" "hdl:tch" "hdl:ltg"
[61] "hdl:glu" "tch:ltg" "tch:glu" "ltg:glu" "y"
```

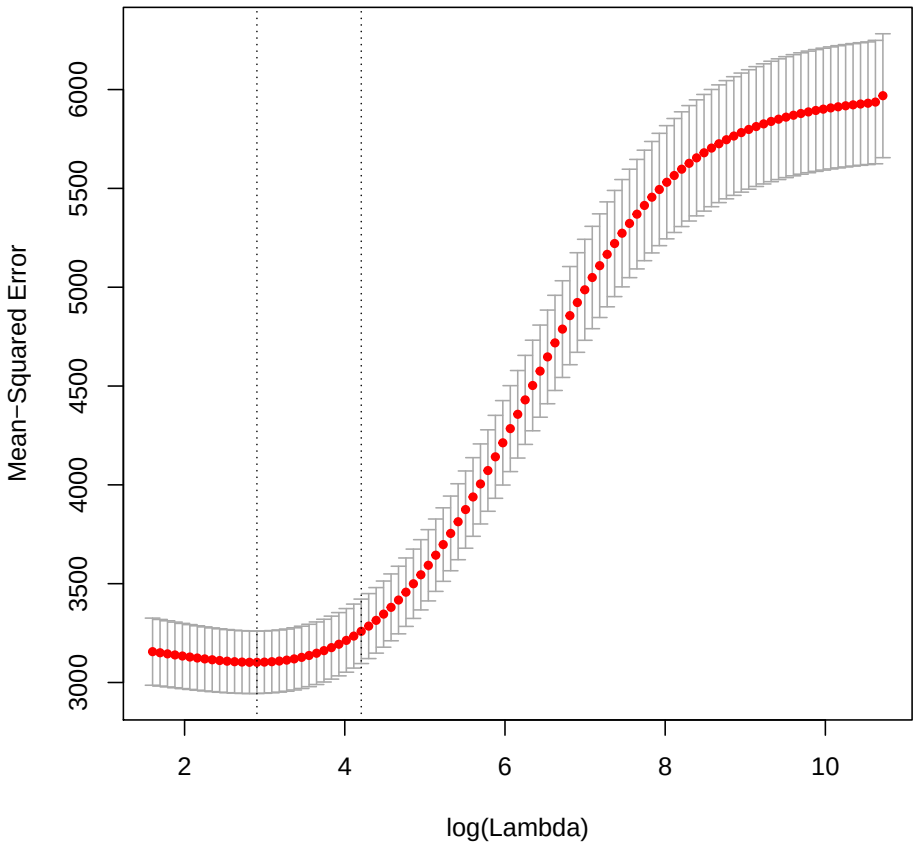
The `glmnet()` function will perform ridge regression when we set `alpha=0`.

```
> library(glmnetUtils)
> ridgefit <- glmnetUtils::glmnet(y ~ ., data=df, alpha=0)
> plot(ridgefit)
```



Cross-validation to tune the shrinkage parameter.

```
> cvridgefit <- glmnetUtils::cv.glmnet(y ~ ., data=df, alpha=0)
> plot(cvridgefit)
```



GLMs

The `glmnet` library (and the `glmnetUtils` wrapper library) allow one to perform ridge regression on generalized linear models.

A penalized maximum likelihood estimate is calculated based on

$$-\lambda \sum_{i=1}^{m} \beta_i^2$$

added to the log-likelihood.

Lasso Regression

Motivation

One drawback of the ridge regression approach is that coefficients will be small, but they will be nonzero.

An alternative appraoch is the **lasso**, which stands for "Least Absolute Shrinkage and Selection Operator".

This performs a similar optimization as ridge, but with an ℓ_1 penalty instead. This changes the geometry of the problem so that coefficients may be zero.

Optimization Goal

Starting with the OLS model assumptions again, we wish to find $\boldsymbol{\beta}$ that minimizes

$$\sum_{j=1}^{n} \left(y_j - \sum_{i=1}^{m} \beta_i x_{ij} \right)^2 + \lambda \sum_{k=1}^{m} |\beta_i|.$$

Note that $\sum_{k=1}^{m} |\beta_i| = \|\boldsymbol{\beta}\|_1$, which is the ℓ_1 vector norm.

As before, the paramater λ is a tuning paramater that controls how much shrinkage and selection occurs.

Solution

There is no closed form solution to this optimization problem, so it must be solved numerically.

Originally, a *quadratic programming* solution was proposed with has $O(n2^m)$ operations.

Then a *least angle regression* solution reduced the solution to $O(nm^2)$ operations.

Modern *coordinate descent* methods have further reduced this to $O(nm)$ operations.

Preprocessing

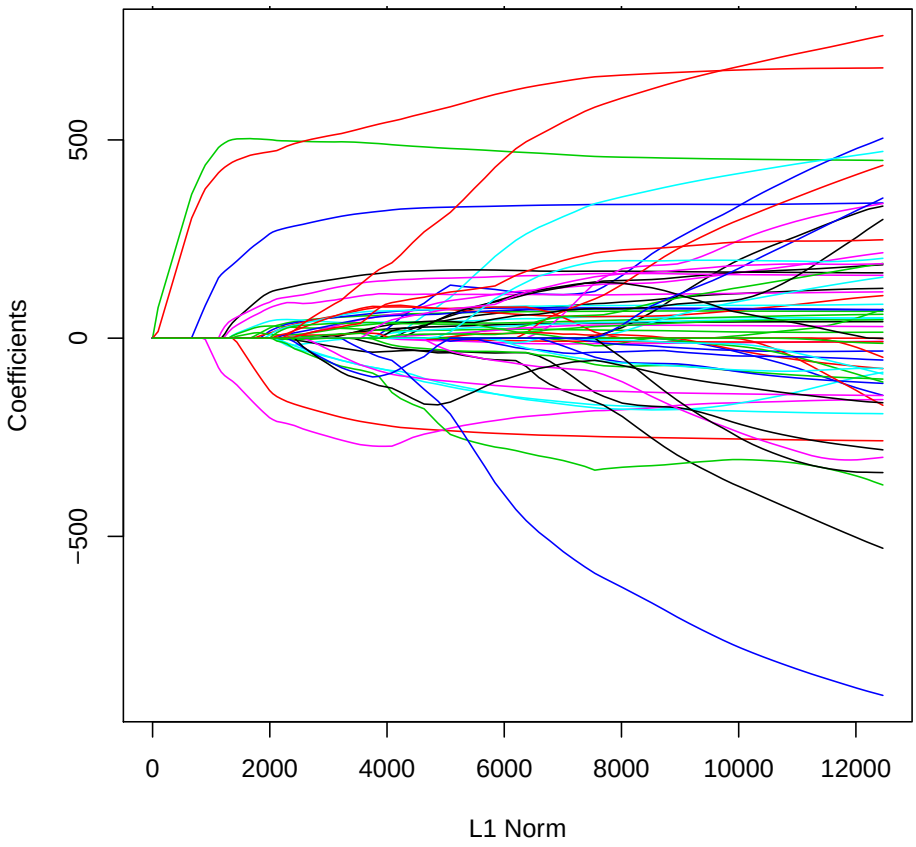
Implicitly...

We mean center $\boldsymbol{y}$.

We also mean center and standard deviation scale each explanatory variable. Why?

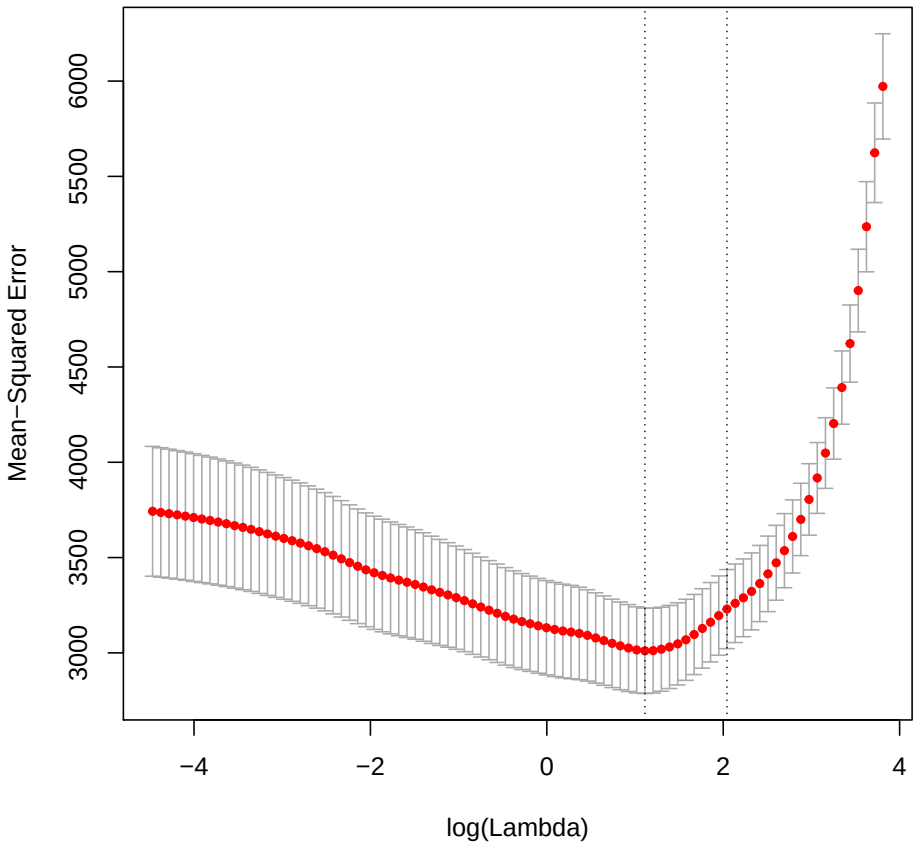
Let's return to the `diabetes` data set. To do lasso regression, we set `alpha=1`.

```
> lassofit <- glmnetUtils::glmnet(y ~ ., data=df, alpha=1)
> plot(lassofit)
```



Cross-validation to tune the shrinkage parameter.

```
> cvlassofit <- glmnetUtils::cv.glmnet(y ~ ., data=df, alpha=1)
> plot(cvlassofit)
```



Bayesian Interpretation

The ridge regression solution is equivalent to maximizing

$$-\frac{1}{2\sigma^2}\sum_{j=1}^{n}\left(y_j-\sum_{i=1}^{m}\beta_i x_{ij}\right)^2-\frac{\lambda}{2\sigma^2}\sum_{k=1}^{m}|\beta_i|$$

which means it can be interpreted as the MAP solution with a Exponential prior on the β_i values.

Inference

Inference on the lasso model fit is difficult. However, there has been recent progress.

One idea proposes a conditional covariance statistic, but this requires all explanatory variables to be uncorrelated.

Another idea called the *knockoff filter* controls the false discovery rate and allows for correlation among explanatory variables.

Both of these ideas have some restrictive assumptions and require the number of observations to exceed the number of explanatory variables, $n > m$.

GLMs

The `glmnet` library (and the `glmnetUtils` wrapper library) allow one to perform lasso regression on generalized linear models.

A penalized maximum likelihood estimate is calculated based on

$$-\lambda \sum_{i=1}^{m} |\beta_i|$$

added to the log-likelihood.

Extras

Source

License

Source Code

Session Information

```
> sessionInfo()
R version 3.3.2 (2016-10-31)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: macOS Sierra 10.12.4

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods
[7] base

other attached packages:
 [1] glmnetUtils_1.0.2 lars_1.2          qvalue_2.1.1
 [4] MASS_7.3-45       broom_0.4.2       dplyr_0.5.0
```

```
 [7] purrr_0.2.2       readr_1.0.0       tidyr_0.6.1
[10] tibble_1.2        ggplot2_2.2.1     tidyverse_1.1.1
[13] knitr_1.15.1      magrittr_1.5      devtools_1.12.0

loaded via a namespace (and not attached):
 [1] reshape2_1.4.2   splines_3.3.2    haven_1.0.0
 [4] lattice_0.20-34  colorspace_1.3-2 htmltools_0.3.5
 [7] yaml_2.1.14      foreign_0.8-67   withr_1.0.2
[10] DBI_0.5-1        modelr_0.1.0     readxl_0.1.1
[13] foreach_1.4.3    plyr_1.8.4       stringr_1.1.0
[16] munsell_0.4.3    gtable_0.2.0     rvest_0.3.2
[19] codetools_0.2-15 psych_1.6.12     memoise_1.0.0
[22] evaluate_0.10    labeling_0.3     forcats_0.2.0
[25] parallel_3.3.2   highr_0.6        Rcpp_0.12.9
[28] scales_0.4.1     backports_1.0.5  jsonlite_1.2
[31] mnormt_1.5-5     hms_0.3          digest_0.6.12
[34] stringi_1.1.2    grid_3.3.2       rprojroot_1.2
[37] tools_3.3.2      lazyeval_0.2.0   glmnet_2.0-5
[40] Matrix_1.2-8     xml2_1.1.1       lubridate_1.6.0
[43] assertthat_0.1   rmarkdown_1.3    httr_1.2.1
[46] iterators_1.0.8  R6_2.2.0         nlme_3.1-131
```