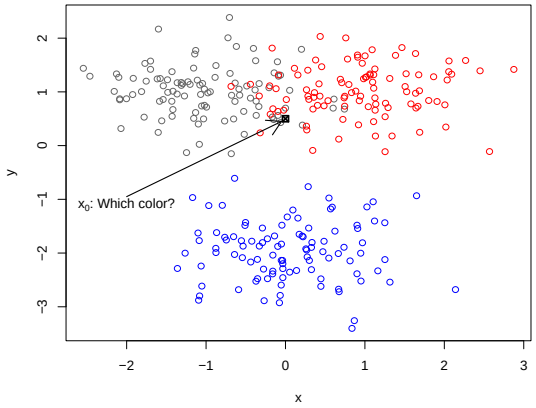
Review: Classification problem

Situation: Have observations $(x_1, y_1), \ldots, (x_n, y_n)$, where $y_i \in \{0, 1, \ldots, J-1\}$ gives a class. Have new observation x_0, want to predict the corresponding class y_0.



Review: Model

We have: $p_j = P(Y = j)$, $f(x|y = j) = f_j(x)$

- $$\pi_j(x_0) = P(Y_0 = j|x_0) = \frac{p_j f_j(x_0)}{\sum_{i=0}^{J-1} p_i f_i(x_0)}$$
- $$ECM(i) = E(c(i|Y)|x_0) = \frac{\sum_{j=0}^{J-1} c(i|j) p_j f_j(x_0)}{\sum_{i=0}^{J-1} p_i f_i(x_0)}$$
- $$\hat{y}_0 = \text{argmin}_i ECM(i) \overset{0/1\text{-loss}}{=} \text{argmax}_i \{p_i f_i(x_0)\}$$

$\Rightarrow$

$$x|y = j \sim \mathcal{N}(\mu_j, \mathbf{\Sigma}_j) \begin{cases} \text{LDA} & \mathbf{\Sigma}_0 = \ldots = \mathbf{\Sigma}_{J-1} \\ \text{QDA} & \text{different } \mathbf{\Sigma}_j \end{cases}$$

Review

We have also discussed:

- k-nearest neighbour algorithm with tuning parameter k.
- Evaluation of classification rules
  - Misclassification rate:

$$P(y_0 \neq \hat{y}(x_0; (x_1, y_1), \ldots, (x_n, y_n)))$$

  - Apparent error rate: (–too optimistic)

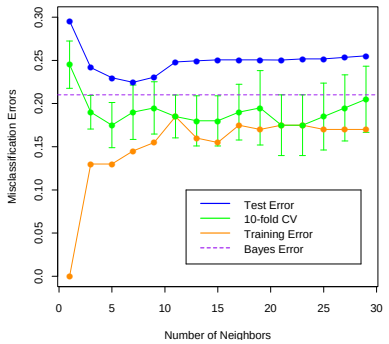
$$\frac{1}{n} \sum_{i=1}^{n} 1(y_i \neq \hat{y}(x_i; (x_1, y_1), \ldots, (x_n, y_n))$$

  - Split training data into a training set and a test set. (–too pessimistic)

Review: Cross-validation (CV)

- Leave-one-out cross validation
- K-fold cross validation, where $K = 5$ or $K = 10$ is often used.

Determining k in knn-algorithm using CV:



Hastie, Tibshirani, Friedman, "The elements of statistical learning", 2nd ed., p. 467

Standard errors for cross validation

See blackboard

Bootstrap

http://tradingconsequences.blogs.edina.ac.uk/files/2013/10/Dr_Martens_black_old.jpg

Bootstrap Bill Turner

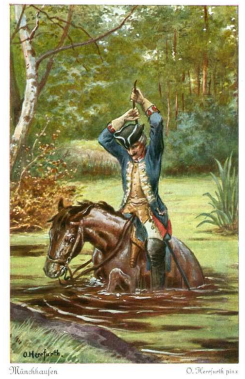
"Bootstrap" Bill Turner from Pirates of the Caribbean.

> *. . . Barbossa tied Bill to a cannon by his bootstraps and sent him to the bottom of the sea.*

http://kidstvmovies.about.com/od/piratesofthecaribbean3/ig/Pirates-At-World-s-End/-Bootstrap-Bill.htm

. . . pull oneself up by one's bootstraps

> *To begin an enterprise or recover from a setback without any outside help; to succeed only on one's own effort or abilities.*
>
> Wiktionary



The term is sometimes attributed to Rudolf Erich Raspe's story "The Surprising Adventures of Baron Munchausen", where the main character pulls himself (and his horse) out of a swamp by his hair

http://redstateeclectic.typepad.com/redstate_commentary/2010/11/sustainability-isnt-sustainable.html

Bootstrapping in statistics

Bootstrap is a computer-based technique for doing statistical inference (usually with a minimum of assumptions). It is not Bayesian.

See blackboard for rough idea

Show animation in R: `boot.iid` in animation package.

Bootstrap principle

Assume we have iid observations from an (unknown) distribution F:

$$F \to (x_1, \ldots, x_n)$$

The empirical distribution function $\hat{F}$ is the CDF that puts mass $1/n$ at each data point x_i:

$$\hat{F}(x) = \frac{1}{n} 1(x_i \leq x)$$

where $1(\cdot)$ denotes the indicator function.

Bootstrap principle

Let θ be an interesting feature of F, $\theta = T(F)$.

For example:

$$\theta = \mathrm{E}(X) = \int x f(x) dx$$

$$\theta = \mathrm{Var}(X) = \int (x - \mathrm{E}(X))^2 f(x) dx$$

The plug-in estimator for θ is defined by:

$$\hat{\theta} = T(\hat{F})$$

The plug-in principle is quite good, if the only information about F, comes from the sample x.

Examples

Thus

$$\theta = \mathrm{E}(X) \Rightarrow \hat{\theta} = \mathrm{E}_{\hat{F}}(X) = \sum_{i=1}^{n} x_i \frac{1}{n} = \bar{x}$$

$$\theta = \mathrm{Var}(X) \Rightarrow \hat{\theta} = \mathrm{Var}_{\hat{F}}(X) = \mathrm{E}_{\hat{F}}[(X - \mu_{\hat{F}})^2]$$

$$= \sum_{i=1}^{n} (x_i - \mu_{\hat{F}})^2 \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$\theta = \mathrm{SD}(X) \Rightarrow \hat{\theta} = \mathrm{SD}_{\hat{F}}(X) = \sqrt{\mathrm{Var}_{\hat{F}}(X)}$$

$$= \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2} \quad \leftarrow \quad \text{standard error}$$

Setting

Assume we have :

$$F \to (x_1, \ldots, x_n)$$

Thus $\hat{F}$ gives mass $\frac{1}{n}$ to each observed value.

A bootstrap sample is defined to be a random sample of size n from $\hat{F}$, say $x^\star = (x_1^\star, \ldots, x_n^\star)$

$$\hat{F} \to (x_1^\star, \ldots, x_n^\star)$$

Simple illustration

Suppose $n = 3$ univariate data points, namely

$$\{x_1, x_2, x_3\} = \{1, 2, 6\}$$

are observed as an iid sample from F that has mean θ . At each observed data value, $\hat{F}$ places mass 1/3. Suppose the estimator to be bootstrapped is the sample mean $\hat{\theta}$.

There are $3^3 = 27$ possible outcomes for $\mathcal{X}^\star = \{X_1^\star, X_2^\star, X_3^\star\}$.

Simple illustration (II)

$\mathcal{X}^\star$	$\hat{\theta}^\star$	$P^\star(\hat{\theta}^\star)$	Observed frequency
1 1 1	3/3	1/27	36/1000
1 1 2	4/3	3/27	101/1000
1 2 2	5/3	3/27	123/1000
2 2 2	6/3	1/27	25/1000
1 1 6	8/3	3/27	104/1000
1 2 6	9/3	6/27	227/1000
2 2 6	10/3	3/27	131/1000
1 6 6	13/3	3/27	111/1000
2 6 6	14/3	3/27	102/1000
6 6 6	18/3	1/27	40/1000

Bootstrap estimate for standard error

- Parameter of interest: $\theta = T(F)$
- Our estimator for θ: $\hat{\theta} = s(x)$
- Want (to estimate) $\mathrm{SD}_F(\hat{\theta})$.

A bootstrap replication of $\hat{\theta}$ is

$$\hat{\theta}^\star = s(x^\star)$$

Use plug-in principle to estimate $\mathrm{SD}_F(\hat{\theta})$. The bootstrap estimate of the standard error of $\hat{\theta} = s(x)$ is $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$. This is called the ideal bootstrap estimate of standard error of $\hat{\theta}$.

Note: Except for very small n, $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$ cannot be computed.
(Number of possible bootstrap sample: n^n.)

Computational way of obtaining a good estimate

We can estimate $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$ by simulation:

1. Generate B bootstrap samples $x^{1\star}, \ldots, x^{B\star}$.
2. Evaluate the corresponding parameter estimates
$$\hat{\theta}^\star(b) = s(x^{b\star}), \quad b = 1, 2, \ldots, B$$
3. Estimate $\mathrm{SD}_{\hat{F}}(\hat{\theta}^\star)$ by
$$\widehat{\mathrm{SE}}_B = \sqrt{\frac{\sum_{b=1}^{B}(\hat{\theta}^\star(b) - \hat{\theta}^\star(\cdot))^2}{B-1}}$$
where
$$\hat{\theta}^\star(\cdot) = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^\star(b)$$

Note
$$\lim_{B\to\infty} \widehat{\mathrm{SE}}_B = \widehat{\mathrm{SE}}_\infty = \widehat{\mathrm{SD}}_{\hat{F}}(\hat{\theta}^\star)$$

Example

Setting
$$\theta = \mathrm{E}(X)$$
$$\hat{\theta} = s(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$
$$\hat{\theta}^\star = s(x^\star) = \frac{1}{n}\sum_{i=1}^{n} x_i^\star = \bar{x^\star}$$

Here, the ideal bootstrap estimate exists

see blackboard

The variability in $\widehat{\mathrm{SE}}_B$ depends on n and B.

How large do we need B?

Intuitively we understand that the $\widehat{\mathrm{SE}}_B$ has larger standard deviation that $\widehat{\mathrm{SE}}_\infty$.

Theory, not to be discussed here, gives the following rules of thumb:

1. Even a small B is informative, say $B = 25$ or $B = 50$ is often enough to get a good estimate of $\mathrm{SE}_F(\hat{\theta})$.
2. Very seldomly more than $B = 200$ is necessary to estimate $\mathrm{SE}_F(\hat{\theta})$.

The parametric bootstrap

When data are modeled to originate from a parametric distribution, so
$$X_1^\star, \ldots, X_n^\star \overset{\text{iid}}{\sim} F(x, \theta)$$
another estimate of F may be employed.

Suppose that the observed data are used to estimate θ by $\hat{\theta}$. Then each parametric bootstrap pseudo-dataset $\mathcal{X}^\star$ can be generated by drawing $X_1^\star, \ldots, X_n^\star \overset{\text{iid}}{\sim} F(x, \hat{\theta})$.

Bootstrapping regression

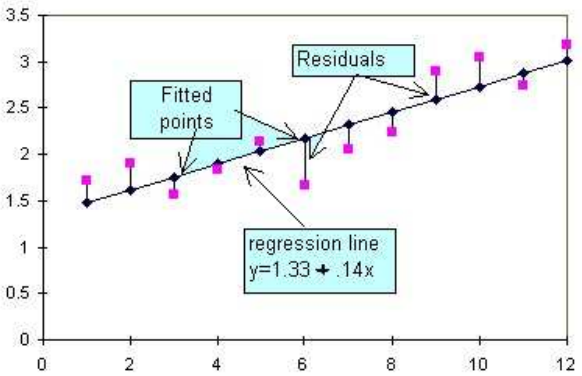
Consider the ordinary multiple regression model

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where ϵ_i are iid mean zero random variables with constant variance.

- Naive: Bootstrapping by resampling from response variables to get distribution of $\hat{\boldsymbol{\beta}}^\star$. However $Y_i|\boldsymbol{x}_i$ are not iid.
- Correct: Bootstrap the residuals.

Review: Residuals



http://fsweb.bainbridge.edu/dbyrd/statistics/regression.htm

Bootstrap the residuals

1. Fit the regression model to the observed data and obtain the fitted responses $\hat{y}_i$ and residuals $\hat{\epsilon}_i$.
2. Sample a bootstrap set of residuals $\hat{\epsilon}_1^\star, \ldots, \hat{\epsilon}_n^\star$ from the set of fitted residuals completely at random and with replacement.
3. Generate a bootstrap set of pseudo responses

$$Y_i^\star = \hat{y}_i + \hat{\epsilon}^\star, \quad \text{for } i = 1, \ldots, n.$$

4. Regress $Y^\star$ on $\boldsymbol{x}$ to obtain a bootstrap estimate $\hat{\boldsymbol{\beta}}^\star$.

Repeat this process to get an empirical distribution of $\hat{\boldsymbol{\beta}}^\star$.

Bootstrapping residuals: Remarks

This approach is also used for autoregressive models, for example.

Bootstrapping the residuals is reliant on

- The models provides an appropriate fitted
- The residuals have a constant variance

Otherwise, a different scheme is recommended.

Paired bootstrap

Suppose response and predictors are measured from a collection of individuals selected at random

$\Rightarrow$ Data pairs $z_i = (x_i, y_i)$ can be regarded as iid realisation from $\boldsymbol{Z}_i = (\boldsymbol{X}_i, Y_i)$ drawn from a joint response-predictor distribution.

Bootstrap:

- Sample $\boldsymbol{Z}_1^\star, \ldots, \boldsymbol{Z}_n^\star$ completely at random with replacement from $z_1, \ldots, z_n$.
- Apply regression model on pseudo dataset to get $\hat{\boldsymbol{\beta}}^\star$.

Repeat this approach many times.

Note: Paired bootstrap is less sensitive to violation of assumptions, e.g. adequacy of regression model, than bootstrapping the residuals.

Copper-nickel alloy

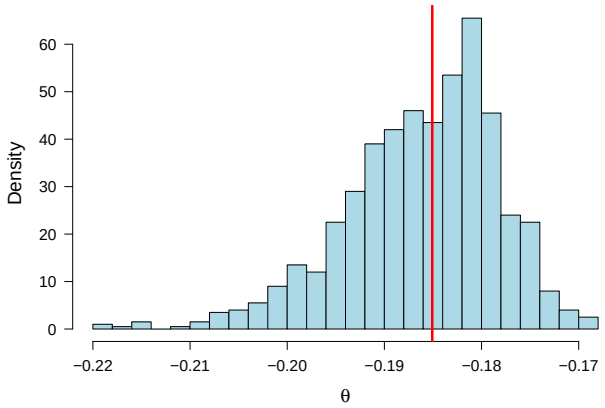
Data: 13 measurements of corrosion loss (y_i) in copper-nickel alloys, each with a specific iron content (x_i).

Question: Change in corrosion loss in the alloys as the iron content increases, relative to corrosion loss where there is no iron, i.e. $\theta = \beta_1/\beta_0$.

x_i	0.01	0.48	0.71	0.95	1.19	0.01	0.48
y_i	127.6	124.0	110.8	103.9	101.5	130.1	122.0
x_i	1.44	0.71	1.96	0.01	1.44	1.96	
y_i	92.3	113.1	83.7	128.0	91.4	86.2	

The observed data yield $\hat{\theta} = \hat{\beta}_1/\hat{\beta}_0 = -0.185$.

Histogram of 10000 bootstrap estimates



Show R-code demo-pairedBootstrap.R

Bootstrap bias correction

The mean value of

$$\hat{\theta}^\star - \hat{\theta}$$

among the pseudo datasets is -0.00125.

The bias-corrected bootstrap estimate of β_1/β_0 is $-0.18507 - (-0.00125) = -0.184$.

Bootstrapping dependent data

Critical requirement: Boostrapped quantities are iid.

Consider a first-order stationary autoregressive process, the AR(1) model:

$$X_t = \alpha X_{t-1} + \epsilon_t$$

where $|\alpha| < 1$ and ϵ_t are iid with mean zero and constant variance.

Here, a method akin to bootstrapping the residuals for linear regression can be applied.

AR(1) model: A model based approach

1. Use a standard method to estimate α
2. Define the estimated innovations $\hat{e}_t = X_t - \hat{\alpha} X_{t-1}$ for $t = 2, \ldots, n$ and let $\bar{\epsilon}$ be the mean of these.
3. Recenter $\hat{e}_t$ to have mean zero by defining $\hat{\epsilon}_t = \hat{e}_t - \bar{e}$.
4. Resample $n + 1$ values from the set $\{\hat{\epsilon}_2, \ldots, \hat{\epsilon}_n\}$ with replacement to yield pseudo innovations $\{\epsilon_0^\star, \ldots, \epsilon_n^\star\}$.
5. Generate pseudo data as $X_0^\star = \epsilon_0^\star$ and $X_t^\star = \hat{\alpha} X_{t-1}^\star + \epsilon_t^\star$ for $t = 1, \ldots, n$.

AR(1) model: A model based approach

Issue: Pseudo-data series is not stationary.

Remedy: Sample larger number of pseudo innovations and generate data series earlier, i.e. $X_k^\star$ for k much less than zero. The first portion of the data can be discarded as burn-in.