SF2930 - Regression analysis

KTH Royal Institute of Technology, Stockholm

Lecture 11 – More about resampling techniques for model assessment
(Iz 5.4, HTF 7.4-7.5, MPV 15.4)

February 18, 2022

Todays lecture

- Test and training errors
- Bootstrap

Test and training errors

Let $\mathcal{T} = \big((\mathbf{x}_i, y_i)\big)_i$ be a training set.

The generalization/prediction/test error

$$\mathsf{Err}_{\mathcal{T}} := \mathbb{E}_{\mathbf{x}_0, y_0}\left[\left|y_0 - \mathbf{x}_0^T \hat{\boldsymbol{\beta}}(\mathcal{T})\right|^2 \mid \mathcal{T}\right]$$

The expected prediction error

$$\mathbb{E}_{\mathcal{T}}[\mathsf{Err}_{\mathcal{T}}]$$

Training error/apparent error rate/in-sample error/regression learning error

$$\mathsf{Err}_{in} := \frac{1}{|\mathcal{T}|} \sum_{i \in T} |y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\mathcal{T}}|^2$$

Regression test error

$$\overline{\mathsf{err}} := \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} |y_i - \mathbf{x}_i^T \hat{\boldsymbol{\beta}}_{\mathcal{T}}|^2$$

Resampling techniques for model assessment

We will now continue developing methods to validate the models we develop using linear regression. In this course, will discuss three such methods:

- Cross validation (random regressors)
- **Bootstrap (random or non-random regressors)**

What is bootstrap?

Motivation

In ideal cases, we have data and a model which is such that all the assumptions we have made earlier, such as normality and independence, hold. However, in many cases, this is not really the case. Also, for some methods, such as e.g. ridge regression, there are no known distributions of parameters which allow us to e.g. calculate confidence intervals.

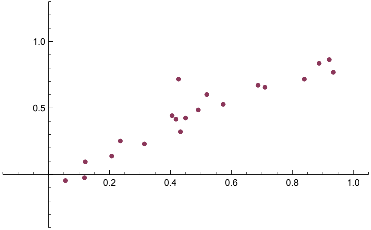
Ideal solution

An ideal solution to this problem would be to consider a lot of independent datasets and then estimate variation, confidence intervals, etc. from their empirical counter-parts. However, we often do not have access to such data.

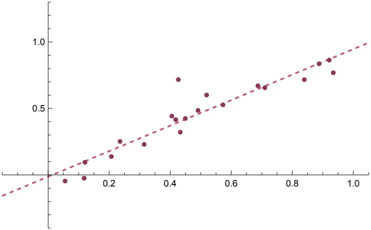
Idea

Use the existing data to "simulate" new data sets, and use these to calculate emperical estimates of the desired parameters.

Bootstrap samples



A sample $\mathcal{T} = \big((\mathbf{x}_i, y_i)\big)_{i\in\{1,2,\dots,n\}}$



A fitted line $y = \beta_0 + \beta_1 x$.

In the next few slides we describe two different methods which can be used to obtain new "samples" from $\mathcal{L}(\mathcal{T})$, known as *bootstrap samples*.

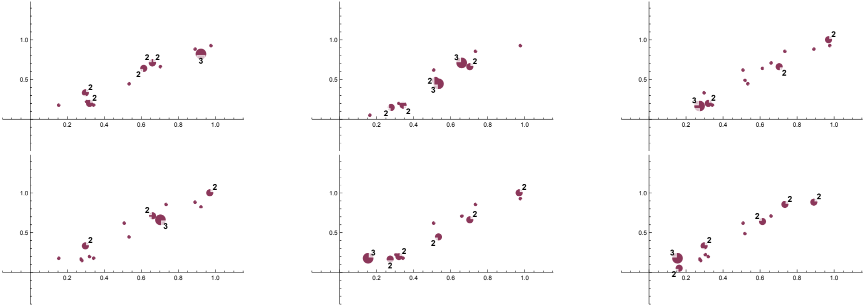
Unconditional/non-parametric bootstrap, bootstrapping cases/pairs

Assumptions

When applying this method, we (ideally) assume that X is random.

Method

For $j = 1, 2, \ldots, m$, pick a bootstrap sample $\mathcal{T}_j^*$ by choosing n observations from $\mathcal{T}$ at random, with replacement.



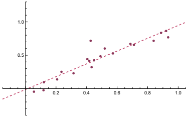
Conditional/parametric bootstrap, bootstrapping residuals

Assumption

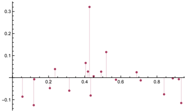
We now assume that the residuals for $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ are i.i.d.

Method

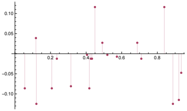
1. Fit a linear regression model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ to obtain an estimate $\boldsymbol{\beta}$ and residuals $\mathbf{e} = \mathbf{y} - X\hat{\boldsymbol{\beta}}$.
2. For $j = 1, 2, \ldots, m$, pick a *bootstrap residual vector* $\mathbf{e}^{*,j}$ by choosing n residuals from $\mathbf{e}$ at random with replacement.
3. Form a *bootstrap vector* of responses by letting $\mathbf{y}^{*,j} := X\hat{\boldsymbol{\beta}} + e^{*,j}$.
4. The *bootstrap sample* $\mathcal{T}^{*,j}$ is given by the pairs $((\mathbf{x}_i, y_i^{*,j}))_{i \in \{1,2,\ldots,n\}}$.



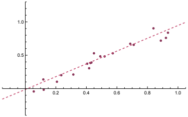
The sample points (x_i, y_i) and the fitted line $\hat{y} = X\hat{\boldsymbol{\beta}}$.



The residuals (x_i, e_i).

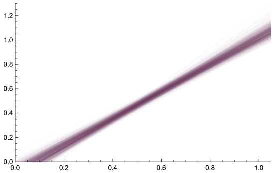


The residuals (x_i, e_i^*).

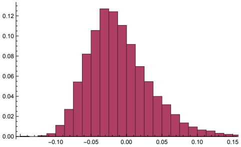


The bootstrap sample points $(x_j, \hat{y}_i + e_i^*)$.

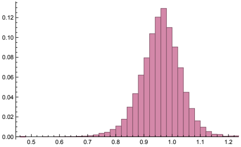
Using bootstrap sample to understand the distribution of $\hat{\beta}$



$y = \hat{\beta}_0((x_i, y_i^{*,j})) + x\hat{\beta}_1((x_i, y_i^{*,j}))$



$\hat{\beta}_0((x_i, y_i^{*,j}))$



$\hat{\beta}_1((x_i, y_i^{*,j}))$

Example

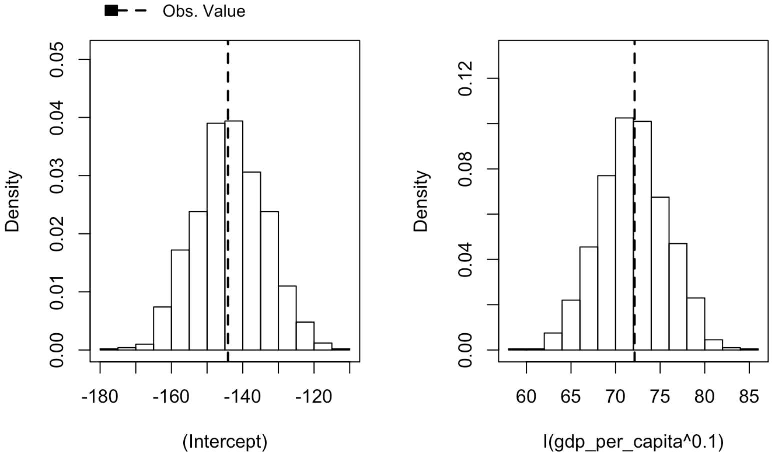
```r
df00.model2 <- lm(people_fully_vaccinated_per_hundred~I(gdp_per_capita^.1), data = df00)
```

```r
library("car")

df00.model2.bootstrapcases <- Boot(df00.model2, 1000, method="case")
df00.model2.bootstrapresiduals <- Boot(df00.model2, 1000, method="residual")
```

Example

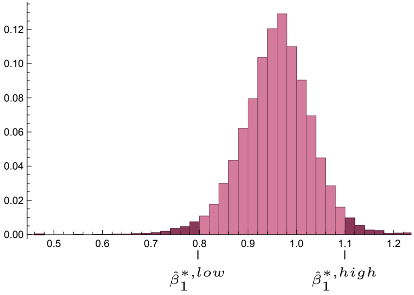
```
hist(df00.model2.bootstrapresiduals, estDensity=FALSE,
     estNormal = FALSE, ci="none")
```



Bootstrap confidence interval for $\hat{\beta}_1$, version 1

Method

1. Using either M1 or M2, obtain bootstrap samples $\mathcal{T}^{*,1}, \mathcal{T}^{*,2}, \ldots, \mathcal{T}^{*,m}$.
2. For each bootstrap sample, calculate $\hat{\boldsymbol{\beta}}^{*,j}$.
3. Let $\mathbb{P}_{\hat{\boldsymbol{\beta}}^*}$ be the empirical distribution of these samples.
4. Let $\hat{\beta}_1^{*,low}$ be the largest number such that $\mathbb{P}_{\hat{\boldsymbol{\beta}}^*}(\hat{\beta}_1^* < \hat{\beta}_1^{*,low}) \leq \alpha/2$, and let $\hat{\beta}_1^{*,high}$ be the smallest number such that $\mathbb{P}_{\hat{\boldsymbol{\beta}}^*}(\hat{\beta}_1^* > \hat{\beta}_1^{*,low}) \leq \alpha/2$.



5. Return: $\hat{\beta}_1^{*,low} \leq \beta \leq \hat{\beta}^{*,high}$.

Example

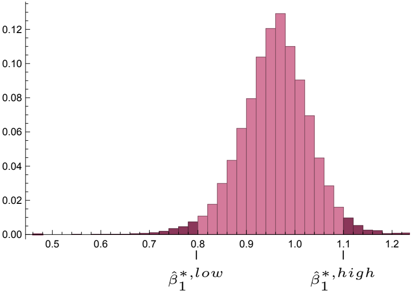
```r
Confint(df00.model2.bootstrapresiduals, level=c(.95, .99),
    type="perc")
```

	Estimate	0.5%	2.5%	97.5%	99.5%
`(Intercept)`	-144.138	-169.000	-164.206	-126.594	-117.838
`I(gdp_per_capita^0.1)`	72.135	62.009	65.174	79.859	81.479

Bootstrap confidence interval for $\hat{\beta}_1$, version 2

Method

1. Using either M1 or M2, obtain bootstrap samples $\mathcal{T}^{*,1}, \mathcal{T}^{*,2}, \ldots, \mathcal{T}^{*,m}$.
2. For each bootstrap sample, calculate $\hat{\boldsymbol{\beta}}^{*,j}$.
3. Let $\mathbb{P}_{\hat{\boldsymbol{\beta}}^*}$ be the empirical distribution of these samples.
4. Let $\hat{\beta}_1^{*,low}$ be the largest number such that $\mathbb{P}_{\hat{\boldsymbol{\beta}}^*}(\hat{\beta}_1^* < \hat{\beta}_1^{*,low}) \leq \alpha/2$, and let $\hat{\beta}_1^{*,high}$ be the smallest number such that $\mathbb{P}_{\hat{\boldsymbol{\beta}}^*}(\hat{\beta}_1^* > \hat{\beta}_1^{*,low}) \leq \alpha/2$.



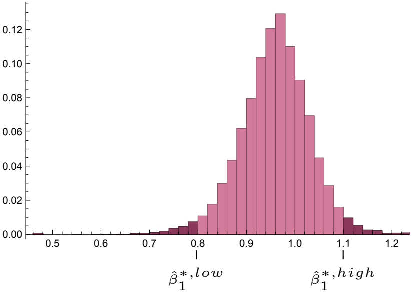
Bootstrap confidence interval for $\hat{\beta}_1$, version 2

5. Note that

$$\hat{\beta}_1^{*,low} \leq \hat{\beta}_1 \leq \hat{\beta}^{*,high} \Leftrightarrow \beta - (\beta - \hat{\beta}_1^{*,low}) \leq \hat{\beta}_1 \leq \beta + (\hat{\beta}^{*,high} - \beta)$$
$$\Leftrightarrow \hat{\beta}_1 - (\hat{\beta}^{*,high} - \beta_1) \leq \beta_1 \leq \hat{\beta}_1 + (\beta_1 - \hat{\beta}_1^{*,low}).$$

If we replace β with $\hat{\beta}_1$ in both of the parentheses above, we obtain the following approximate $100(1-\alpha)\%$ confidence interval for β_1:

$$\hat{\beta}_1 - (\hat{\beta}^{*,high} - \hat{\beta}_1) \leq \beta \leq \hat{\beta}_1 + (\hat{\beta}_1 - \hat{\beta}_1^{*,low}).$$



Example

```r
Confint(df00.model2.bootstrapresiduals, level=c(.95, .99),
    type="bca")
```

	Estimate	0.5%	2.5%	97.5%	99.5%
(Intercept)	-144.138	-170.763	-165.413	-125.587	-118.974
I(gdp_per_capita^0.1)	72.135	61.508	64.729	80.243	82.445

Example

```
Confint(df00.model2.bootstrapresiduals, level=c(.95, .99),
    type="bca")
```

	Estimate	0.5%	2.5%	97.5%	99.5%
(Intercept)	-144.138	-170.763	-165.413	-125.587	-118.974
I(gdp_per_capita^0.1)	72.135	61.508	64.729	80.243	82.445

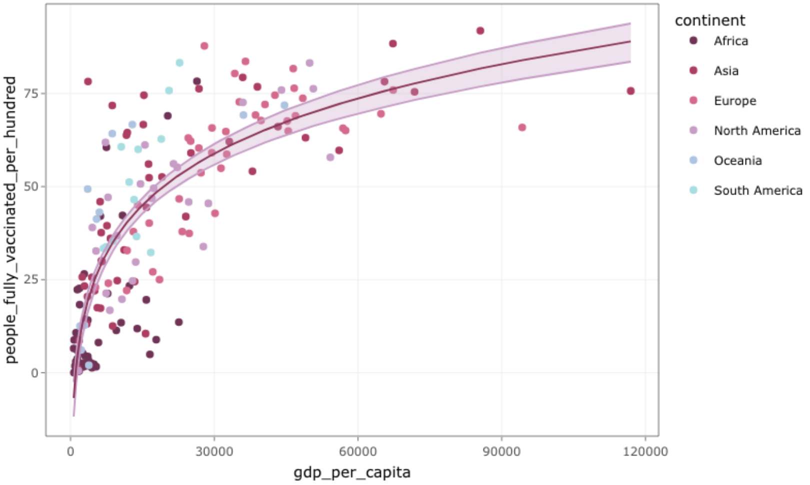
```
Confint(df00.model2.bootstrapresiduals, level=c(.95, .99),
    type="perc")
```

	Estimate	0.5%	2.5%	97.5%	99.5%
(Intercept)	-144.138	-169.000	-164.206	-126.594	-117.838
I(gdp_per_capita^0.1)	72.135	62.009	65.174	79.859	81.479

```
confint(df00.model2)
```

	Estimate	2.5%	97.5%
(Intercept)	-144.138	-64.51172	-124.63379
I(gdp_per_capita^0.1)	72.135	62.009	79.75718

Example


continent
Africa
Asia
Europe
North America
Oceania
South America
people_fully_vaccinated_per_hundred
0
25
50
75
gdp_per_capita
0
30000
60000
90000
120000

Simple bootstrap estimate for the prediction error

1. Obtain bootstrap samples $\mathcal{T}^{*,1}, \mathcal{T}^{*,2}, \dots, \mathcal{T}^{*,m}$.
2. For each bootstrap sample, calculate
$$\widehat{PE}_i = \frac{1}{n}\|\mathbf{y} - X\hat{\boldsymbol{\beta}}^{*,j}\|_2^2$$
3. Take an average over all samples to obtain an estimate of the prediction error,
$$\widehat{PE} = \frac{1}{m}\sum_{j=1}^{m}\widehat{PE}_j.$$

$\widehat{PE}$ is sometimes called the *simple bootstrap estimate* of the prediction error, or the bootstrap estimate of the training error $\overline{\mathrm{err}}$.

Comments

- The simple bootstrap error will in general be overly optimistic, since we the bootstrap samples will have data points in common with the original sample. There are versions of this procedure which are better.

Example

```
all_fits <- as.matrix(df00.model2.bootstrapcases$t[,1]) %*% t(as.matrix(rep(1, nrow(df00))))+as.matrix(df00.model2.bootstrapcases$t[,2]) %*% t(as.matrix(df00$gdp_per_capita^.1))

responses <- as.matrix(rep(1, nrow(all_fits))) %*% t(as.matrix(df00$people_fully_vaccinated_per_hundred))

mean((all_fits-responses)^2)
```

```
[1] 249.1058
```

The apparent error rate

1. Obtain bootstrap samples $\mathcal{T}^{*,1}$, $\mathcal{T}^{*,2}$, $\dots$, $\mathcal{T}^{*,m}$.
2. For each bootstrap sample, calculate

$$\widehat{PE}_j = \frac{1}{n}\|\mathbf{y}^{*,j} - X^{*,j}\hat{\boldsymbol{\beta}}^{*,j})\|_2^2.$$

3. Take an average over all samples to obtain an estimate of the prediction error,

$$\widehat{PE} = \frac{1}{m}\sum_{j=1}^{m}\widehat{PE}_j.$$

In this case, $\widehat{PE}$ is known as the *apparent error rate*, which is a bootstrap estimate of the in-sample error Err_{in}.

Example

Bootstrapping cases

```r
library("boot")

df00.fit <- function(data) {
  mod <- lm(people_fully_vaccinated_per_hundred~I(gdp_per_capita^.1), data = data)

  mean(mod$residuals^2)
}

case.fun <- function(d,i)
  df00.fit(d[i,])

df00.case <- boot(df00, case.fun, R=999)

mean(df00.case$t)
```

```
[1] 244.0289
```

Example

Bootstrapping residuals

```r
library("boot")

df00.fit <- function(data) {
  mod <- lm(people_fully_vaccinated_per_hundred~I(gdp_per_capita^.1), data = data)

  mean(mod$residuals^2)}

df00$fit <- fitted(df00.model2)
df00$res <- resid(df00.model2)

model.fun <- function(d,i) {
  d$people_fully_vaccinated_per_hundred <- d$fit+d$res[i]
  df00.fit(d) }

df00.mod <- boot(df00, model.fun,  R=999)

mean(df00.mod$t)
```

```
[1] 244.0455
```

Example

Bootstrapping cases

```r
library("boot")

df00.fit <- function(data) {
  mod <- lm(people_fully_vaccinated_per_hundred~I(gdp_per_capita^.1), data = data)

  c(coef(mod),sqrt(summary(mod)$coef[,2]^2)) # first coefficients, then standard errors  }

case.fun <- function(d,i) { df00.fit(d[i,]) }

df00.case <- boot(df00, case.fun, R=999)

summary(df00.case)
# confint(df00.case, level=.99, type="bca")
```

```
Number of bootstrap replications R = 999

  original  bootBias   bootSE   bootMed
1 -144.138 -0.168079  8.1083 -144.604
2   72.134  0.032791  3.2204   72.197
3    9.886 -0.025054  0.8305    9.8375
4    3.864 -0.010819  0.3196    3.8375
```

Example

Comments

- `bootBias` $= \overline{\hat{\beta}_j^*} - \hat{\beta}_j$
- `bootSE` $= \sqrt{\frac{1}{n-1} \sum_i (\hat{\beta}_j^{*,i} - \overline{\hat{\beta}_j^*})^2}$

It possible to get the same table, in special cases, from the output from `Boot`:

```
summary(df00.model2.bootstrapresiduals)
```

```
Number of bootstrap replications R = 1000

                     original  bootBias bootSE  bootMed
(Intercept)          -144.138   0.31934 9.9842 -144.02
I(gdp_per_capita^0.1)  72.134 -0.13545 3.9201   72.15
```

Compare with the output from `lm` (uses normality assumption!)

```
summary(df00.model2)
```

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -144.138      9.886  -14.58   <2e-16 ***
I(gdp_per_capita^0.1) 72.134        3.864   18.67   <2e-16 ***
```

When can you not apply bootstrap?

Random vs not random

Non-constant or dependent errors