MSG400-TMS150

Stochastic data processing and simulation 2020/21

Bootstrap

1 Introduction

In the previous two lectures we have focussed on the selection of appropriate linear regression models, their parameters estimation and the assessment of the estimates uncertainty. However, the estimation of parameters in general models and the assessment of the estimates' variability, is not a problem exclusively pertaining regression models (linear or nonlinear). In fact, the estimation of parameters in probability distributions is a central problem in statistics that one tends to encounter already during the very first course on the subject. More generally, we are interested in inferences for *unknowns*: these are not just model parameters, for example when we predicted new observations, clearly those are unobserved quantities. Along with the estimate of some unobserved quantity we are (we should be!) also interested in its accuracy, which can be described in terms of the bias and the variance of the estimator, as well as confidence intervals around it. Sometimes, such measures of accuracy can be derived analytically. Often, they can not. The *bootstrap* is a technique that can be used to estimate them numerically from a single data set[1].

2 The general idea

Let $X_1, ..., X_n$ be a i.i.d. sample from distribution F, that is $\mathbf{P}(X_i \leq x) = F(x)$, and let $X_{(1)}, ..., X_{(n)}$ be the corresponding ordered sample. For the purpose of this introduction, suppose we are interested in some *scalar* parameter θ which is associated with this distribution (mean, median, variance etc), the treatment of a multivariate θ being analogous. There is also an estimator $\hat{\theta} = t(\{X_1, ..., X_n\})$, with t denoting some function, that we can use to estimate θ from data. In this setting, it is the deviation of $\hat{\theta}$ from θ that is most interesting.

Ideally, to get an approximation of the estimator distribution we would like to repeat the data-generating experiment, say, B times, calculating $\hat{\theta}$ for each of the B data sets. That is, we would draw B samples of size n from the true distribution F (with replacement, if F is discrete). In practice, this is impossible.

The bootstrap method is based on the following simple idea: *Even if we do not know F we can approximate it from data and use this approximation, $\hat{F}$, instead of F itself.*

This idea leads to several flavours of bootstrap that deviate in how, exactly, the approximation $\hat{F}$ is obtained. Two broad areas are the *parametric* and *non-parametric* bootstrap.

[1] More generally, bootstrap is a method of approximating the distribution of functions of the data (statistics), which can serve different purposes, among others the construction of CI and hypothesis testing.

The non-parametric estimate is the so called empirical distribution (you will see the corresponding pdf if you simply do a histogram of the data), that can be formally defined as follows:

Definition 2.1. *With # denoting the number of members of a set, the* empirical distribution function $\hat{F}$ *is given by*

$$\hat{F}(x) = \frac{\#\{i : X_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{X_i \leq x\},$$

$$\mathbb{I}\{X_i \leq x\} = \begin{cases} 1 & \text{if } X_i \leq x \\ 0 & \text{otherwise} \end{cases}$$

That is, it is a discrete distribution that puts mass $1/n$ on each data point in your sample.

The parametric estimate assumes that the data comes from a certain distribution family (Normal, Gamma etc). That is, we say that we know the general functional form of the pdf, but not the exact parameters. Those parameters can then be estimated from the data (typically with Maximum Likelihood) and plugged in the pdf to get $\hat{F}$. This estimation method leads to more accurate inference if we guessed the distribution family correctly but, on the other hand, $\hat{F}$ may be quite far from F if the family assumption is wrong.

2.1 Algorithms

The two algorithms below describe how the bootstrap can be implemented.

Non-parametric bootstrap
Assuming a data set $x = (x_1, ..., x_n)$ is available.

1. Fix the number of bootstrap re-samples B. Often $B \in [1000, 2000]$.
2. Sample a new data set x^* set of size n from x *with replacement* (this is equivalent to sampling from the empirical cdf $\hat{F}$).
3. Estimate θ from x^*. Call the estimate $\hat{\theta}_1^*$. Store.
4. Repeat step 2 and 3 further $B - 1$ times.
5. Consider the empirical distribution of $(\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$ as an approximation of the true distribution of $\hat{\theta}$.

You may produce the histogram of $(\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$. This histogram represents the pdf of $\hat{F}$.

Parametric bootstrap
Assuming a data set $x = (x_1, ..., x_n)$ is available.

1. Assume that the data comes from a known distribution family F_ψ described by a set of parameters ψ (for a Normal distribution $\psi = (\mu, \sigma)$ with μ being the expected value and σ the standard deviation).
2. Estimate ψ with, for example, Maximum likelihood, obtaining the estimate $\hat{\psi}$.
3. Fix the number of bootstrap samples B. Often $B \in [1000, 2000]$.
4. Sample a new data set x^* set of size n from $F_{\hat{\psi}}$.
5. Estimate θ from x^*. Call the estimate $\hat{\theta}_1^*$. Store.
6. Repeat 4 and 5 further $B - 1$ times.
7. Consider the empirical distribution of $(\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$ as an approximation of the true distribution of $\hat{\theta}$.

Again, you may produce the histogram of $(\hat{\theta}_1^*, ..., \hat{\theta}_B^*)$, this representing the pdf of $\hat{F}$.

Concretely, let us say that $X_i \sim N(0, 1)$, θ is the median and it is estimated by $\hat{\theta} = X_{(n/2)}$, the $n/2$-th element in the ordered sequence. In Figure 1 the distribution of $\hat{\theta}$ approximated with the non-parametric and parametric bootstrap is plotted. Note that the parametric distribution is smoother than the non-parametric one, since the samples were drawn from a continuous distribution.

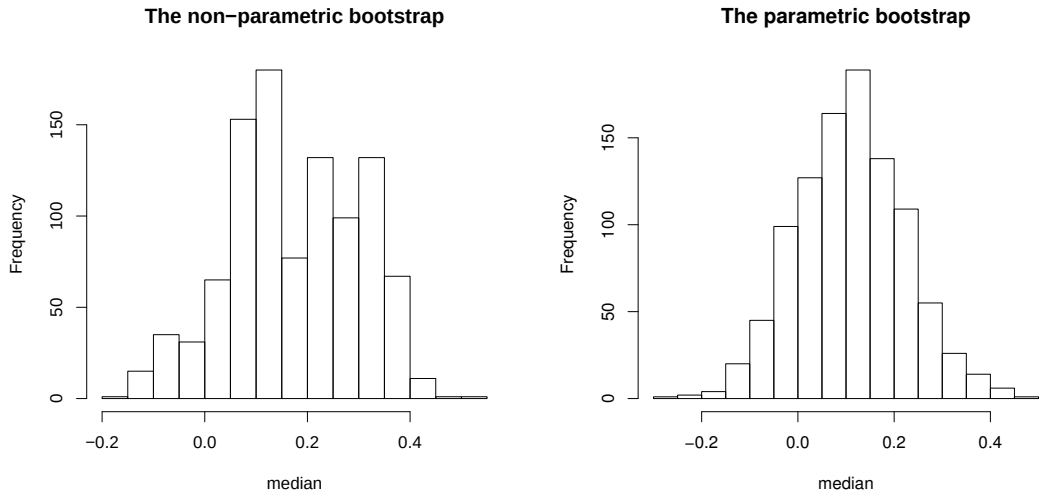

Figure 1: The non-parametric and the parametric bootstrap distribution of the median, $B = 1000$.

3 Bias and variance estimation

The theoretical bias and variance of an estimator $\hat{\theta}$ are defined as

$$\mathbf{Bias}(\hat{\theta}) = \mathbf{E}[\hat{\theta} - \theta] = \mathbf{E}[\hat{\theta}] - \theta$$

$$\mathbf{Var}(\hat{\theta}) = \mathbf{E}\left[(\hat{\theta} - \mathbf{E}[\hat{\theta}])^2\right]$$

In words, the bias is a measure of a systematic error ($\hat{\theta}$ tends to be either smaller of larger than θ) while the variance is a measure of random error.

In order to obtain the bootstrap estimates of bias and variance we plug in the original estimate $\hat{\theta}$ (which is a constant given data) in place of θ and $\hat{\theta}^*$ (the distribution of which we get from bootstrap) in place of $\hat{\theta}$. This leads us to the following approximations:

$$\mathbf{Bias}(\hat{\theta}) \approx \frac{1}{B}\sum_i \hat{\theta}_i^* - \hat{\theta} = \bar{\hat{\theta}}^* - \hat{\theta}$$

$$\mathbf{Var}(\hat{\theta}) \approx \frac{1}{B-1}\sum_i (\hat{\theta}_i^* - \bar{\hat{\theta}}^*)^2$$

That is, the variance is, as usual, estimated by the sample variance (but for the bootstrap sample of $\hat{\theta}$) and bias is estimated by how much the original $\hat{\theta}$ deviates from the average of the bootstrap sample denoted $\bar{\hat{\theta}}^*$.

4 Confidence intervals

There are several methods for CI construction with Bootstrap, the most popular being "normal", "basic" and "percentile". Let $\hat{\theta}^*_{(i)}$ be the ordered bootstrap estimates, with $i = 1, ..., B$ indicating the different samples. Let α be the significance level. In all that follows, $\hat{\theta}^*_\alpha$ will denote the α-quantile of the distribution of $\hat{\theta}^*$. You can approximate this quantile with $\hat{\theta}^*_{([(B+1)\alpha])}$ with the $[x]$ indicating some interpolation or rounding procedure[2].

Basic	$[2\hat{\theta} - \hat{\theta}^*_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}^*_{\alpha/2}]$
Normal	$[\hat{\theta} - z_{1-\alpha/2}\hat{se}, \hat{\theta} - z_{\alpha/2}\hat{se}]$
Percentile	$[\hat{\theta}^*_{\alpha/2}, \hat{\theta}^*_{1-\alpha/2}]$

with z_α denoting an α quantile from a Normal distribution and $\hat{se}$ the estimated standard deviation of $\hat{\theta}$ calculated from the bootstrap sample.

Basic CI

To obtain this confidence interval we start with $W = \hat{\theta} - \theta$ (compare to the classic CI that correspond to a t-test). If the distribution of W was known, then a two-sided CI could be obtained by considering $\mathbf{P}(w_{\alpha/2} \leq W \leq w_{1-\alpha/2}) = 1 - \alpha$, which leads to CI $= [l_{low}, l_{up}] = [\hat{\theta} - w_{1-\alpha/2}, \hat{\theta} - w_{\alpha/2}]$. However, the distribution of W is not known, and is approximated with the distribution for $W^* = \hat{\theta}^* - \hat{\theta}$, with w^*_α denoting the corresponding α-quantile. The CI

[2]There is no general agreement on the rounding scheme to use. Some sources use ceiling (integer closest to x from above) and others use flooring (integer closest to x from below).

becomes $[\hat{\theta} - w^*_{1-\alpha/2}, \hat{\theta} - w^*_{\alpha/2}]$. Noting that $w^*_\alpha = \hat{\theta}^*_\alpha - \hat{\theta}$ and substituting this expression in the CI formulation leads to the definition given in the box.

This interval construction relies on an assumption about the distribution of $\hat{\theta} - \theta$, namely that it is independent of θ. This assumption, called pivotality, does not necessarily hold in most cases. However, the interval gives acceptable results even if $\hat{\theta} - \theta$ is close to pivotal.

Normal CI

This confidence interval probably looks familiar since it is almost an exact replica of the commonly used confidence interval for a population mean. Indeed, similarly to that familiar CI, we can get it if we have reasons to believe that $Z = (\hat{\theta} - \theta)/\hat{se} \sim N(0, 1)$ (often true asymptotically as the data size $n \to \infty$). Alternatively, we can again consider $W = \hat{\theta} - \theta$, place the restriction that it should be normal and estimate its variance with bootstrap. Observe that the normal CI also implicitly assumes pivotality.

Percentile CI

These type of CI may seem natural and obvious but actually requires a quite convoluted argument to motivate. Without going into details, you get this CI by assuming that there exists a transformation h such that the distribution of $h(\hat{\theta}) - h(\theta)$ is pivotal, symmetric and centered around 0. You then construct a Basic CI for $h(\theta)$ rather than θ itself, get the α quantiles and transform those back to the original scale by applying h^{-1}. Observe that although we do not need to know h explicitly, the existence of such a transformation is a must, which is not always the case.

5 Limitations of bootstrap

Yes, those do exist. Bootstrap tends to give results easily (too much so, in fact), but it is possible that those results are completely wrong. More than that, they can be completely wrong without being obvious about it. The following are some such situations.

5.1 Infinite variance

In the "general idea" of bootstrapping we plugged in an estimate $\hat{F}$ of F, and then sampled from it. This works only if $\hat{F}$ actually is a good estimate of F, that is it captures the essential features of F despite being based on a finite sample. This may not be the case if F is very heavy tailed, i.e. has infinite variance. The intuition is that in this case extremely large or small values can occur, and when they do they have a great effect on the bootstrap estimate of the distribution of $\hat{\theta}$, making it unstable. As a consequence, the measures of accuracy of $\hat{\theta}$, such as CI, will be unreliable.

The classic example of this is the mean estimator $\hat{\theta} = \bar{X}$ with the data generated from a Cauchy distribution. For it, both the first and the second moments (e.g. mean and variance) are infinite, leading to nonsensical confidence intervals even for large sample sizes. A less obvious example is a non-central Student t distribution with 2 degrees of freedom. This distribution has pdf

$$f_X(x) = \frac{1}{2\,(1 + (x - \mu)^2)^{3/2}} \quad \text{for} \quad x \in \mathbb{R}.$$

where μ is the location parameter, defined as $\mathbf{E}[X]$. So, the first moment is finite and its estimator $\bar{X}$ is consistent. The second moment, however, is infinite, and the right tail of the

distribution grows heavier with increasing μ. This leads to the 95% CI coverage probabilities that are quite far from the supposed 95% even when the sample size n is as large as 500 (that is the percentage of confidence intervals that are supposed to contain the true parameter value would be far from the expected 95%).

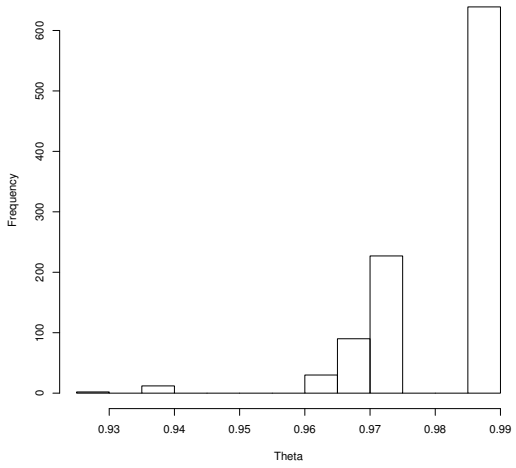

Figure 2: Histogram of 1000 non-parametric bootstrap samples of $\hat{\theta}^*$, the data from the uniform $U(0, \theta)$ distribution. $\theta = 1$.

5.2 Parameter on the boundary

The classical example here is the $U(0, \theta)$ distribution. The maximum likelihood estimate $\hat{\theta}$ is simply $\max(x_1, ..., x_n)$, which will always be biased ($\hat{\theta} < \theta$). In this case the non-parametric bootstrap leads to a very discrete distribution, with more than half of the bootstrap estimates $\hat{\theta}^*$ equal to the original $\hat{\theta}$ (Figure 2). Clearly, if the quantiles used in CIs are taken from this distribution the results will be far from accurate. However, the parametric bootstrap will give a much smoother distribution and more reliable results.

5.3 Lack of pivotality

In all the CI descriptions above the word "pivotality" shows up. So we can guess that it is a bad thing not to have. To circumvent this, something called "studentized bootstrap" can be used.

The idea behind the method is simple and can be seen as an extrapolation of the basic bootstrap CI. There, we looked at $W = \hat{\theta} - \theta$. Now, we instead consider the standardized version $W = (\hat{\theta} - \theta)/\sigma$, with σ denoting the standard deviation of $\hat{\theta}$. The confidence interval will then be calculated through

$$\mathbf{P}(w_{\alpha/2} \leq W \leq w_{1-\alpha/2}) = \mathbf{P}(\hat{\theta} - w_{1-\alpha/2}\sigma \leq \theta \leq \hat{\theta} - w_{\alpha/2}\sigma) = 1 - \alpha$$

As with the basic CI, the distribution of W is not known and has to be approximated, this time with the distribution of $W^* = (\hat{\theta}^* - \hat{\theta})/\hat{se}^*$. Here, $\hat{se}^*$ denotes the standard deviation corresponding to *each bootstrap sample*.

This CI is considered to be more reliable than the ones described earlier. However, there is a catch, and this catch is the estimate of the standard deviation of $\hat{\theta}^*$, $\hat{se}^*$. This estimate is not easily obtained. You can get it either parametrically (but then you need a model which you probably don't have) or through re-sampling. This means that you have to do an extra bootstrap for each of the original bootstrap samples. That is, you will have a loop within a loop, and it can become very heavy computationally.

5.4 Further reading

For a wider and deeper treatment of the bootstrap see chapters 10 and 11 in *Computer Age Statistical Inference* by Bradley Efron and Trevor Hastie. The PDF has been made freely available from the publisher at `https://web.stanford.edu/~hastie/CASI/`.

6 Simulation of pseudo-random numbers

Strictly speaking, this topic is not part of the bootstrap. However, the ability to simulate pseudo-random numbers is often necessary whenever we wish to produce simulations from a stochastic model, or draw samples from a probability distribution. Simulation of special pseudo-random numbers is, in fact, required in the exercises that follow.

Generating samples from an arbitrary distribution can be a complex task. For challenging problems you may be interested in taking the very interesting course MVE187/MSA101 "Computational methods for Bayesian statistics". For very simple cases, the so-called *inverse transform method* will work, when the distribution function F is analytically tractable.

Theorem 6.1 (Inverse transform method)**.** *Let $U \sim U(0,1)$ be a uniform random variable. Consider a random variable X having distribution function F, i.e. we write $X \sim F$. Then we have that $F^{-1}(U) \sim F$, that is the random variable $F^{-1}(U)$ is distributed as F.*

Proof: $Pr(F^{-1}(U) \leq x) = Pr(U \leq F(x)) = F(x)$, which means that the distribution function of $F^{-1}(U)$ is F.

The theorem above is instrumental to easily sample from one-dimensional or two dimensional distributions, again when F^{-1} is tractable. In fact, we just need to sample a single $u \sim U(0,1)$ to then obtain a draw $x^* \sim F$ where $x^* = F^{-1}(u)$, and x^* is the draw we wanted. Clearly we have that $X = F^{-1}(U)$ and hence $U = F(X)$.

Example 6.1. *We write a procedure to sample from an exponential distribution $X \sim \text{Exp}(\lambda)$ with mean $\lambda > 0$. Notice we use the definition of distribution function below (other versions replace $1/\lambda$ with λ)*

$$F(x) = 1 - e^{-\frac{x}{\lambda}}.$$

Above we have written that $U = F(X)$ which means that we can equate $u = F(x) = 1 - e^{-\frac{x}{\lambda}}$. After some trivial algebra we obtain $x = -\lambda \ln(1-u)$, which suffices to generate exponential draws with mean λ. Below is a MATLAB code generating 10,000 independent exponential draws with mean $\lambda = 3$, and Figure 3 shows that their (normalized) histogram matches the theoretical probability density function (PDF) of the distribution very well. Notice, the normalization of the histogram means that its area is approximately equal to 1, just like the integral of a PDF. Without normalization the comparison would be meaningless.

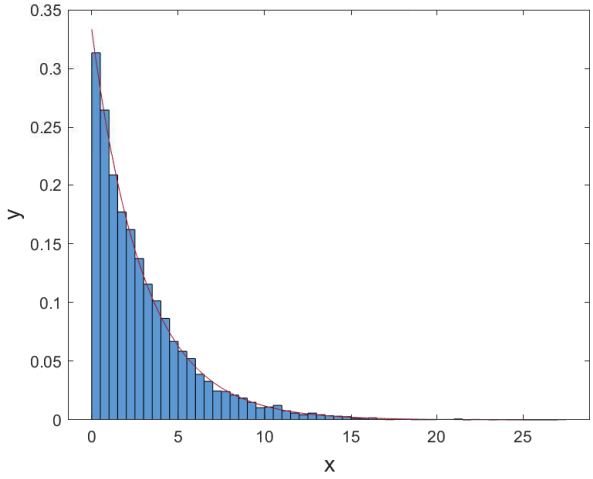

Figure 3: Normalized histogram of 10,000 exponential draws with mean 3 and comparison with the theoretical PDF $y = 1/\lambda \cdot \exp(-x/\lambda)$.

```
rng(123) % set a seed, for reproducibility
lambda = 3;
u = rand(10000,1); % 10000 uniforms in (0,1)
x = -lambda*log(1-u);  % exponential draws
histogram(x,'Normalization','pdf')  % normalized histogram (so that it is comparable with a pdf)
hold on  % useful to add the next plot to the current one
xgrid = [0:.01:25];
ypdf = 1/lambda * exp(-xgrid/lambda);
plot(xgrid,ypdf)
xlabel('x','FontSize',14)  % add x label and set font size
ylabel('y','FontSize',14)  % add y label and set font size
```

Notice that, unlike in R, the function to produce a vector of uniform random numbers expects two dimensions to be provided. That is `rand(10000,1)` generates a 10000×1 vector. We could have also tried `rand(1,10000)` and it would have been fine. But it would be a mistake to use `rand(10000)`, as this would create a $10,000 \times 10,000$ matrix. For info on MATLAB functions use `help`, e.g. `help rand` in this case.

Assignment A2: second part

Here follow exercises for the second part of A2. See `model-choice.pdf` for the first part. Notice that **one exercise has to be solved in Matlab and the other one with R.**

Exercise 3 (to be solved in MATLAB), 4.5 points

"The significant-wave-height is the average height of the highest one-third of all measured waves, which is equivalent to the estimate that would be made by a visual observer at sea."

The data file `atlantic.txt` contains the significant-wave-height recorded 14 times a month during several winter months in the north Atlantic. It was found that a good fit to the empirical distribution of the data is given by a Gumbel distribution. This is a distribution for rare events. For example, it is useful in predicting the chance that an extreme earthquake, flood or other natural disaster will occur. It has distribution function

$$F(x;\mu,\beta) = \exp\Big(-\exp\big(-\frac{x-\mu}{\beta}\big)\Big), \qquad x \in \mathbb{R}$$

for "location" parameter $\mu \in \mathbb{R}$ and "scale" parameter $\beta > 0$ to be determined. You can estimate the parameters from the data by maximum likelihood using the provided Matlab function `est_gumbel.m`.

(a) [0.5 points] Using the inversion cdf method, derive the formula to generate random draws from a Gumbel distribution (write the maths in detail).

(b) [0.5 points] Using the formula derived in (a) together with the maximum likelihood estimates obtained from the atlantic data, simulate a sample of Gumbel data of size n (where n is the same as for the `atlantic` data), then using a qqplot check that the distributions of the `atlantic` data and the simulated data approximately agree.

(c) [1 point] Provide parametric bootstrapped 95% confidence intervals for the parameters, using the percentile method, based on $B = 10,000$ simulations. [Please place `rng(123)` before the for-loop to ease grading.]

The expected 100-year return value of the significant-wave-height gives the largest expected value during a 100-year period. The Tth return value is given by $F^{-1}(1-1/T;\mu,\beta)$. We note that we have 14 observations during a month and three winter months during a year, thus for a 100-year return $T = 3 \cdot 14 \cdot 100$.

(d) [1 point] Provide a parametric bootstrapped 95% confidence interval for the 100-year return value, using the percentile method. [tip: you may reuse the bootstrapped parameters obtained in (c)].

(e) [0.5 points] Based on (d), supposing for a moment that the `atlantic.txt` values have been collected close to a coast, if you were to advise the city council where measurements have been taken, how tall *at a minimum* should a barrier be in order to protect the city from the highest (expected) waves?

(f) [1 point] Above we used parametric bootstrapping. Suppose now to be in the situation where simulating artificial data from the Gumbel model is computationally very expensive (this is true for many realistic simulators of natural phenomena). When model simulation is expensive, it is useful to consider the nonparameric bootstrap instead. Repeat (c), except that here you use the nonparametric bootstrap. What do you conclude? [Please place `rng(123)` before the for-loop to ease grading.]

Exercise 4 (to be solved in R), 1.5 points

Reprise the Wage data from A2 part 1. In that exercise it was of course evident that simple linear regression would be inappropriate for the data. Now, we want to compute some approximate confidence intervals for the parameters of the polynomial of order $p = 3$. This is interesting because in this course we have not introduced the theory for how to compute exact confidence intervals for the parameters of general linear models, we have only considered the case of simple linear regression (1 covariate). Therefore approximate solutions provided via bootstrap is a pedagogically useful example of what we can do when we have no clue about the existence of exact solutions.

Here we pretend to know the following and **nothing else**: (i) we know how to use R to obtain the estimates of a polynomial of any order, and (ii) we know how to sample from the available data with replacement, but that we do not know how to simulate observations from a linear regression model (this is incorrect, but let's just pretend this). Set a simulation that produces 90% bootstrap confidence intervals for each parameter of the order $p = 3$ polynomial, based on $B = 2000$ bootstrap samples. Compare with the results returned from applying `confint` to such model. [tip: to sample data you can use `sample.int` while remembering to activate the `replace` option]. Please place `set.seed(321)` before the for-loop.

Assignment A2 has a maximum of 11 points: the exercises above represent the second part of A2. The first part of A2 assigns max 5 points. The second part assigns max 6 points. The entire A2 assignment (part 1 + part 2) must be written as a single LaTeX report and be submitted to Canvas. That is you should not write two separate reports. Notice the "recommended deadline" on the course webpage.

Please use the report template provided on the course page, and in case you worked out the exercises with some student, remember to write the name of said student as a footnote in the report front page. Also, recall that each submitted report and the code therein is an INDIVIDUAL submission, not group work.

Finally, since you have to write a proper report: as from the provided template, you are also asked to produce some background on the methodology you use. So do not just write answers to the exercise questions.

For a given project report, 0.5 points will be deducted if the report is not clearly structured or is otherwise hard to understand. Likewise, 0.5 points will be deducted if the code attached to the report is not properly structured and commented. The report should not be longer than 10 pages including figures, but excluding appendix. Figures and axes labels should be big enough to be readable if printed. It is OK to use colors.

Full details on grading are of course available at the course webpage.