Data Science and Statistical Computing

Assignment 1 Solutions

Q1

Write down the formula to compute x_i given (w_i, γ_i) for day i. Hence, or otherwise, write down a formula for the statistic $S(\mathbf{w}, \boldsymbol{\gamma})$ which computes the mean wind power generation rate, given vectors of wind speed and angle of incidence for n days, $\mathbf{w} = (w_1, \ldots, w_n), \boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$.

$$
\begin{aligned}
x_i &= w_i \cos \gamma_i \\
\implies S(\mathbf{w}, \boldsymbol{\gamma}) &= \frac{1}{n} \sum_{i=1}^{n} V(w_i \cos \gamma_i)
\end{aligned}
$$

where $V(\cdot)$ is as given in the question.

Q2

Compute the mean power generation rate, $S(\mathbf{w}^\star, \boldsymbol{\gamma}^\star)$, for the bootstrap resample:

$$(\mathbf{w}^\star, \boldsymbol{\gamma}^\star) = \big((w_2, \gamma_2), (w_3, \gamma_3), (w_1, \gamma_1), (w_2, \gamma_2)\big)$$

Also, what is the largest value $S(\mathbf{w}^\star, \boldsymbol{\gamma}^\star)$ could take for any possible bootstrap resample?

NOTE/HINT: resampling is done for *days*, since there is some unknown joint distribution for (w_i, γ_i) ... hence the same index must always apply to the individual resamples of wind and angle or else the joint distribution is violated (eg (w_2, γ_2) is valid, but (w_2, γ_4) is *not* valid)

We extend the table in the question to compute the power for each observation:

Day i	w_i	γ_i		x_i		$V(x_i)$
1	6.9	-54	$6.9\cos(-54)$	$= 4.056$	$V(4.056) = 200(4.056) - 200$	$= 611.2$
2	6.0	72	$6.0\cos(72)$	$= 1.854$	$V(1.854) = 200(1.854) - 200$	$= 170.8$
3	4.5	80	$4.5\cos(80)$	$= 0.781$	$V(0.781) = 0$	$= 0$
4	1.9	29	$1.9\cos(29)$	$= 1.662$	$V(1.662) = 200(1.662) - 200$	$= 132.4$

$$
\begin{aligned}
S(\mathbf{w}^\star, \boldsymbol{\gamma}^\star) &= \frac{1}{4}\Big(2V\big(6\cos 72\big) + V\big(4.5\cos 80\big) + V\big(6.9\cos(-54)\big)\Big) \\
&= \frac{1}{4}\Big(2\big(170.8\big) + \big(0\big) + \big(611.2\big)\Big) \\
&= 238.2
\end{aligned}
$$

The largest value $S(\mathbf{w}^\star, \boldsymbol{\gamma}^\star)$ could take is when,

$$(\mathbf{w}^\star, \boldsymbol{\gamma}^\star) = ((w_1, \gamma_1), (w_1, \gamma_1), (w_1, \gamma_1), (w_1, \gamma_1))$$

when we would have $S(\mathbf{w}^\star, \boldsymbol{\gamma}^\star) = 611.2$.

Q3

> Bootstrap simulation for the data above gives $\bar{S}^\star = 228.56$ and $\widehat{\text{Var}}\big(S(\mathbf{w}, \boldsymbol{\gamma})\big) = 13350$. Compute the 95% Normal confidence interval (CI) and comment on whether you consider there to be any appreciable bias in the mean estimator.

The 95% Normal confidence interval is given by:

$$\hat{\theta} \pm z_{0.025}\sqrt{\widehat{\text{Var}}(S(\mathbf{x}))}$$

Here, $\hat{\theta} = \frac{1}{4}(611.2 + 170.8 + 0 + 132.4) = 228.6$, so the CI is

$$228.6 \pm 1.96\sqrt{13350} = 228.6 \pm 226.5 = [2.1, 455.1]$$

NOTE: the confidence interval should be centred on $\hat{\theta} = S(\mathbf{x})$ and *not* on $\bar{S}^\star$, although in this particular case it would make almost no difference.

In this setting,

$$\widehat{\text{bias}}(\theta, \hat{\theta}) = \bar{S}^\star - \hat{\theta} = 228.56 - 228.6 = -0.04$$

The estimated bias, -0.04, should be looked at in relation to the scale of the estimator, 228.6. Here, we would clearly attribute such a small bias on the estimator scale to sampling variation and conclude this is most likely an unbiased estimator.

Q4

> Following are sorted bootstrapped values for $S(\mathbf{w}^\star, \boldsymbol{\gamma}^\star)$ and how often they occured ('Freq'). The total number of bootstrap replicates (sum of 'Freq') is $B = 1000$. Find a 95% CI for the mean power generation rate by the percentile method. Comment on the resulting CI compared to the Normal interval.

The percentile methods means we are seeking the values at the empirical 2.5% and 97.5% quantiles. Thus for 1000 resamples, for the 25th and 975th ordered values. From the tables,

$$5 + 15 = 20 + 21 = 41$$
$$3 + 14 = 17 + 16 = 33$$

Thus, the 25th ordered value lies within the repeats of 42.7 and the 975th in the repeats of 491.4, giving the confidence interval $[42.7, 491.4]$.

There is an appreciable difference in the Normal and percentile confidence intervals, so we would tend to think that perhaps the assumption that the distribution of the statistic had reached Normality might be incorrect and prefer the percentile interval.

Q5

> The power company believe that the component x, which contributes to power generation, can be directly modelled as Exponential, with parameter λ depending on conditions near the turbine:
>
> Exponential pdf: $\quad f(x \mid \lambda) = \lambda e^{-\lambda x}, \quad x \in [0, \infty), \lambda > 0$
>
> In other words, with the Exponential distribution as a model for x_i, we can simulate it directly, without simulating (w_i, γ_i).
>
> Making this assumption, list the detailed steps to perform a parametric bootstrap estimate of the uncertainty in the mean power generation rate (kW) ('detailed' means you should include any derivations of, for example, maximum likelihood estimators or any other quantities needed to do parametric bootstrap)

Step 1

We require the maximum likelihood estimator (MLE) for the parameter λ in an Exponential distribution, noting that the data are the univariate perpendicular wind speeds, x_i, and not (w_i, γ_i).

$$\begin{aligned}
\ell(\lambda; \mathbf{x}) &= \prod_{i=1}^{n} \lambda e^{-\lambda x_i} \\
\implies \mathcal{L}(\lambda; \mathbf{x}) &= \sum_{i=1}^{n} \log \lambda - \lambda x_i \\
&= n \log \lambda - \lambda \sum_{i=1}^{n} x_i \\
\implies \frac{\partial \mathcal{L}}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \\
\therefore \quad \frac{n}{\hat{\lambda}} - \sum_{i=1}^{n} x_i &= 0 \\
\implies \hat{\lambda} &= \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}
\end{aligned}$$

Therefore, for the MLE we have that,

$$\begin{aligned}
\bar{x} &= \frac{4.056 + 1.854 + 0.781 + 1.662}{4} = 2.088 \\
\implies \hat{\lambda} &= 0.4789
\end{aligned}$$

Step 2

Simulate many data sets of size 4, $(x_1^{\star b}, x_2^{\star b}, x_3^{\star b}, x_4^{\star b}), b = 1, \ldots, B$, with each $x_i^\star$ simulated from an Exponential$(\lambda = 0.4789)$.

Each time compute the statistic of interest,

$$S_b^\star = \frac{1}{4} \sum_{i=1}^{4} V(x_i^{\star b})$$

to obtain another bootstrap estimate of the mean wind power generation rate.

Step 3

Compute $\bar{S}^\star = \frac{1}{B} \sum_{b=1}^{B} S_b^\star$ and $\widehat{\text{Var}}(S(\cdot))$ in the usual way based on the many bootstrap estimated mean wind power generation rates.

Q6

> Code up your detailed steps from Q5 in R and use $B = 10,000$ parametric bootstrap replicates to estimate $\widehat{\text{Var}}\big(S(\cdot)\big)$. You may find the function `rexp()` useful (see help file). Provide both the R code as well as the estimate of $\widehat{\text{Var}}\big(S(\cdot)\big)$ it produces in your answer (*do not* list all 10000 simulations!!)
>
> **NOTE:** Do not worry, the solutions used for marking will allow a range of values in the answer for $\widehat{\text{Var}}\big(S(\cdot)\big)$ to allow for variability due to the random simulation.

```r
library("dplyr")

# Information from the question
# Data:
wind <- data.frame(
  w = c(6.9, 6.0, 4.5, 1.9),
  gamma = c(-54, 72, 80, 29)
)
# Power curve:
V <- function(x) {
  ifelse(x < 1,
         0,
         ifelse(x < 12,
                200*x - 200,
                2200))
}

# Compute perpendicular component and power
wind <- wind |>
  mutate(x = w*cos(gamma*pi/180),
         V = V(x))
wind
```

```
    w gamma         x        V
1 6.9   -54 4.0557182 611.1436
2 6.0    72 1.8541020 170.8204
3 4.5    80 0.7814168   0.0000
4 1.9    29 1.6617774 132.3555
```

```r
# MLE for Exponential model
lambda.hat <- 1/mean(wind$x)
lambda.hat
```

```
[1] 0.478869
```

```r
# Number of bootstraps
B <- 10000

# Statistic of interest
S <- function(x) {
  mean(V(x))
}

# Perform bootstrap
S.star <- rep(0, B)
for(b in 1:B) {
  x.star <- rexp(4, lambda.hat)
  S.star[b] <- S(x.star)
}

# Estimator
S(wind$x)
```

```
[1] 228.5799
```

```r
# Variance of estimator
var(S.star)
```

```
[1] 35279.64
```

This is a lot higher than the variance produced for the non-parametric bootstrap estimator.

When producing these solutions, I ran the above estimator with $B = 10000$ many times to account for the stochastic nature of the estimate. The minimum and maximum on those runs gives a plausible solution range of $[33000, 39000]$ ... any answer for $\widehat{\text{Var}}\big(S(\cdot)\big)$ in this range is an acceptable solution.