Problem Sheet 4

MATH50011
Statistical Modelling 1

Week 4

Lecture 7 (Proof of MLE Consistency and Asymptotic Normality)

1. In the lecture notes, we saw that MLEs are asymptotically normal and sketched a proof of this (subject to regularity conditions). Many other estimators are also the solutions to estimating equations.

   Let $X_1, \ldots, X_n$ be i.i.d. real-valued random variables and suppose that we wish to estimate the value of $\theta_0 \in \mathbb{R}$ defined as the unique $E[\psi(X_1, \theta)] = 0$ for a twice continuously differentiable function $\psi : \mathbb{R}^2 \to \mathbb{R}$. Define $\hat{\theta}_n$ as the unique solution to $\sum_{i=1}^n \psi(X_i, \theta) = 0$.

   Sketch a proof that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, \sigma^2(\theta_0))$ and find an expression for $\sigma^2(\theta_0)$. At what steps in your proof sketch do you need additional assumptions required to justify an operation?

2. In the notation of Problem Sheet 3.10 (the R lab), define the one-step estimator

   $$\hat{\theta}_n^{(1)} = T_n + I_n(T_n)^{-1} U_n(T_n).$$

   Suppose that T_n is asymptotically normal and that $I_n(T_n)$ is consistent for the Fisher information $I_f(\theta)$. Show that

   $$\sqrt{n}(\hat{\theta}_n^{(1)} - \theta_0) \to_d N(0, I_f(\theta_0)^{-1}).$$

   Hint: use a first-order Taylor expansion of $U_n(\theta)$.

Lecture 8 (Confidence Intervals)

3. Dr. Jetson asked a random sample of 10000 UK households whether or not they own a robotic vacuum cleaner. She finds that 1300 of the households own a robotic vacuum and the other 8700 do not. Based on this data, she estimates that 13% of UK households own a robotic vacuum with a 95% confidence interval of 12.3% to 13.7%. Dr. Jetson tells you that

   > "There is a 95% probability that between 12.3% and 13.7% of UK households own a robotic vacuum cleaner."

   What is the main problem with the above statement? Provide a correct description of the confidence interval suitable for a non-statistician.

4. A random sample of 11 components in a factory is collected. The length in cm of each component is recorded below

   3.26 1.76 1.63 1.79 2.43 0.88 0.99 1.12 4.56 2.11 2.73

   Assume that the lengths are normally distributed with mean μ and variance σ^2. Construct a 99% confidence interval for μ.

5. Let $Y_1, \dots, Y_n$ be i.i.d. Exp(λ), where $\lambda > 0$ is unknown.

   (a) Show that $2\lambda \sum_{i=1}^{n} Y_i$ has a χ^2-distribution with $2n$ degrees of freedom;

   (b) Derive a $(1 - \alpha) \times 100\%$ confidence interval for λ;

   (c) Using the following observations, compute a 95% confidence interval for λ.

   1.04 1.39 0.1 2.04 4.73 0.89 0.51 0.89 0.66 0.93

   (Note: for $X \sim \chi^2_{20}$, $P(X \leq 9.59) = 0.025$ and $P(X \leq 34.17) = 0.975$.)

6. Find an approximate 95% confidence interval for the odds that a randomly selected UK household owns a robotic vacuum based on the data in Exercise 3. (Hint: use the delta method.)

7. Use the Bonferroni correction to find a 95% confidence region for (μ, σ^2) based on a random sample $X_1, \dots, X_n$ from a $N(\mu, \sigma^2)$ distribution. Apply your result to construct a 95% confidence region for (μ, σ^2) based on the data in Exercise 4.

R lab: The Bootstrap

This exercise introduces concepts through use of the R software package.

Let T_n be an asymptotically normal estimator of θ based on a random sample $Y_1, \ldots, Y_n$. We now consider a flexible method called the bootstrap that allows us to approximate the sampling distribution of T_n using observations $y_1, \ldots, y_n$.

The bootstrap sampling distribution can be used to construct confidence intervals for θ by either:

i. Computing $\mathrm{SE}(T_n)$ with respect to the bootstrap sampling distribution and using the formula $T_n \pm c_{\alpha/2}\,\mathrm{SE}(T_n)$ from the notes;

ii. Computing the $\alpha/2$ and $1 - \alpha/2$ quantiles of the bootstrap sampling distribution.

This procedure is widely applicable, but is most useful for estimators where it is difficult to obtain a closed-form expression for $\mathrm{SE}(T_n)$.

In R, the code below shows how we usually compute $\bar{y}$ and estimate its standard error based on 4 data points: 2, 4, 9, and 12.

```
y <- c(2,4,9,12)
ybar <- mean(y)
se.ybar <- sqrt(var(y)/4)
```

Running the above code, we find that the standard error is about 2.29.

The bootstrap sampling distribution of $\bar{Y}$ is obtained by resampling the data points with replacement and computing $\bar{Y}$ based on the resampled data. There are 4 data points, so there are $4^4 = 256$ equally likely resamples. We can use R to obtain all 256 values in the bootstrap sampling distribution as follows:

```
# All 4^4 = 256 possible resamples with replacement
y.star <- expand.grid(y,y,y,y)

# All 256 sample means based on resampling w/replacement
ybar.star <- apply(y.star, 1, mean)

# The standard error based on this is
se.ybar.star <- sqrt(var(ybar.star))
```

From the above, we find that the bootstrap standard error is about 1.98.

It is a good idea to also visualise the bootstrap distribution of $\bar{Y}$. This can be achieve with hist(ybar.star). For large sample sizes, we would expect the bootstrap sampling distribution to look approximately normal. The normal approximation for $n = 4$ seems to be less than ideal.

The number of bootstrap resamples n^n grows too quickly to be reasonable for the average statistician. Instead, we usually approximate the bootstrap sampling distribution by drawing a large number of random samples as follows.

```
set.seed(50011)

ybar.boot <- numeric(length = 10000)
for(i in 1:10000){
y.boot <- sample(y, size = 4, replace = TRUE)
ybar.boot[i] <- mean(y.boot)
}
se.ybar.boot <- sqrt(var(ybar.boot))
```

From the above, we find that the bootstrap standard error is about 1.99. This is nearly the same as the value obtained by enumerating all 256 samples.

8. Using the code examples above:

   (a) Construct three approximate 95% confidence intervals for the mean μ based on the formula $T_n \pm c_{\alpha/2}\,\mathrm{SE}(T_n)$ where the standard error is based on se.ybar, se.ybar.star, and se.ybar.boot.

   (b) Construct two additional approximate 95% confidence intervals for the mean with limits define by the 2.5% and 97.5% percentiles of ybar.star and ybar.boot. (Hint: use the quantile() function.)

   (c) Compare the similarities/differences in the confidence intervals you constructed in parts (a) and (b).

   (d) Replace the data y in your code with a random sample of $n = 30$ standard exponential random variables: y <- rexp(n=30). Based on your previous code, construct an approximate 95% confidence interval for the mean based on a normal approximation. Construct two different bootstrap 95% confidence intervals for the mean based on 10000 resamples. (Note: you are not being asked to enumerate all 30^{30} resamples.)