Computer class 3 exercises

Richard Wilkinson

Question 1

- Generate 100 $N(0,1)$ random variables. Plot the CDF and overlay the true $N(0,1)$ CDF on top.
- Estimate the probability that $X < 0.5$ using the ECDF and calculate the true value. Note that in R, `ecdf(X)` returns the ECDF of a random sample X as a function. This can then be evaluated, for example, `ecdf(X)(0.5)` etc
- Repeat this process a large number of times to convince yourself that your estimator is unbiased.
- [*Optional*] The Dvoretzky-Kiefer-Wolfowith (DKW) inequality says that for any $\epsilon > 0$, and any $n > 0$, then for all $x \in \mathbb{R}$

$$P(\sup_x |F_n(x) - F(x)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$$

  where F_n is the ECDF based on a sample of size n, and F is the true CDF. Convince yourself empirically that this result is true.

Question 2

Consider the `hills` dataset in the MASS package in R, which contains data on the record time for each of 35 Scottish hill races. We want to build a model to predict the record time on the basis of the race distance and the total amount of height gained during the route. Because we are worried about outliers in the data, we will use a robust regression approach using M estimators.

We can fit a robust linear model to this dataset using the command

```
library(MASS)
fit <- rlm(time~dist+climb, hills, maxit=100)
summary(fit)
```

```
##
## Call: rlm(formula = time ~ dist + climb, data = hills, maxit = 100)
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10.75039  -3.28395  -0.03358   3.53791  65.70100
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept) -9.6067  1.7545    -5.4754
## dist         6.5507  0.2451    26.7237
## climb        0.0083  0.0008     9.9199
##
## Residual standard error: 5.209 on 32 degrees of freedom
```

The coefficient standard errors reported by `rlm` rely on asymptotic approximations, and may not be trustworthy in a sample of size 35. Thus, we will use bootstrapping to estimate confidence intervals.

1. Calculate a bootstrap 95% confidence interval for the coefficient of `dist` using model-based resampling.
2. Recalculate this confidence interval by now using case resampling - i.e. by bootstrap resampling (x_i, y_i), re-fitting the model to each bootstrap sample, and forming the bootstrap distribution of β. Contrast

this with your answer from the previous part, and with a 95% confidence interval obtained from the asymptotic standard error estimates reported by the `summary(fit)` command.

3. An alternative model is proposed, which includes an interaction term between `dist` and `climb` and a quadratic `climb` term. Calculate the mean-square prediction error for both models using leave-one-out cross validation. Which model do you prefer?

```
fit2 <- rlm(time~dist*climb+I(climb^2), hills, maxit=100)
```

Note that there are built-in commands in R for doing bootstrapping and cross-validation. You should *not* use these (except to check your answer), but instead write your own R commands.

Question 3

Failure of the bootstrap: Suppose $X_1, \ldots, X_n \sim U[0, \theta]$.

1. Show that the maximum likelihood estimator of θ is

$$\hat{\theta} = \max_{i=1,\ldots,n} X_i$$

2. Calculate the CDF of $\hat{\theta}$

3. Generate a toy dataset of size $n = 100$ (assuming that $\theta = 1$) using the code

```
set.seed(1)
n=100
X = runif(n,min=0, max=1)
```

4. The parametric bootstrap works by generating a bootstrap sample from the fitted parametric model, i.e., simulating

$$X_1^*, \ldots, X_n^* \sim U[0, \hat{\theta}],$$

and then fitting the parametric model, i.e., setting

$$\hat{\theta}^* = \max_{i=1,\ldots,n} X_i^*$$

Explain why for the parametric bootstrap

$$\mathbb{P}(\hat{\theta}^* = \hat{\theta}) = 0.$$

Simulate 10^5 (parametric) bootstrap replicates and compare the distribution of these with the non-parametric bootstrap estimate.

5. The non-parametric bootstrap creates bootstrap samples by sampling from the empirical CDF. Simulate 10^5 non-parametric bootstrap samples and compare the true distribution of $\hat{\theta}$ to the histogram from the non-parametric bootstrap and with the parametric bootstrap.

6. If $\hat{\theta}^*$ is a bootstrap estimate of $\hat{\theta}$, show that

$$\mathbb{P}(\hat{\theta}^* = \hat{\theta}) = 1 - (1 - (1/n))^n$$

and hence that

$$\mathbb{P}(\hat{\theta}^* = \hat{\theta}) \to 1 - e^{-1} \approx 0.632$$

as $n \to \infty$.