OLLSCOIL NA hÉIREANN, CORCAIGH
THE NATIONAL UNIVERSITY OF IRELAND, CORK

COLÁISTE NA hOLLSCOILE, CORCAIGH
UNIVERSITY COLLEGE, CORK

Examination Session and Year	Winter 2021
Module Code	ST4060 ST6015 ST6040
Module Name	Statistical Methods for Machine Learning I Computer Analytical Techniques for Actuarial Applications Machine Learning and Statistical Analytics I
Paper Number	Paper Number: 1
External Examiner	Mr. Andrew Maclaren
Head of Department	Dr. Kevin Hayes
Internal Examiner(s)	Dr. Eric Wolsztynski
Instructions to Candidates	• Please answer all questions. • Provide all your answers in the Word document provided. • Paste your R code into the Word document at the end of each question. • Submit a pdf version of your final Word document for Canvas submission.
Duration of Paper	3 hours.

List of required R libraries:

```
ISLR
```

List of (possibly) useful R functions:

```
apply()
approx()
as.numeric()
boxplot()
cbind()
coef()
colnames()
fitted()
lines()
lm()
matrix()
mean()
median()
na.omit()
nls()
nrow()
numeric()
order()
par()
plot()
points()
predict()
quantile()
sample()
sd()
set.seed()
smooth.spline()
sqrt()
sum()
summary()
which()
```

Question 1 [15 marks]

No code is required for this question.

Let $S_M(X)$ denote a cubic spline evaluated at some d-dimensional design point $\mathbf{X} \in \Omega \subset \mathbb{R}^d$, and for a set of values $\{\xi_i\}_{i=1}^M \in \Omega$, with $\alpha_k \in \mathbb{R}$, $\beta_k \in \mathbb{R}$, $\forall k$:

$$S_M(X) = \sum_{k=0}^{3} \beta_k X^k + \sum_{j=1}^{M} \alpha_j (X - \xi_j)_+^3$$

using notation $(u)_+ = \max(u, 0)$. Let us now define the following criterion $C_\lambda(g)$ as a function of some continuous function $g(X)$, with the aim to fit a sample of N observations $\{Y_i\}_{i=1}^N$, using some parameter $\lambda \in \mathbb{R}^+$:

$$C_\lambda(g) = \sum_{i=1}^{N} (Y_i - g(X_i))^2 + \lambda \int_\Omega \left(g''(u)\right)^2 du$$

(a) Name the function estimate $\hat{g}$ that minimizes the above criterion $C_\lambda(g)$.

(b) Using the information given above, provide a name for the function $\hat{g}$ that minimizes $C_\lambda(g)$, and indicate how this function should be evaluated given the data (X, Y) for this problem.

(c) Describe, in a few sentences, the effect of parameter λ on the estimate $\hat{g}$.

Solution:

(a) This is a penalized spline.

(b) The solution to this optimisation problem is a cubic spline, i.e. $\hat{g}(X) = S_M(X)$, evaluated at the design points $\xi_i = X_i$.

(c) A larger value of λ will place more emphasis on the penalty term, which will gradually become more important over the standard Least Squares terms. Since this penalty evaluates the magnitude of the second-order derivative of the spline $\hat{g}$, a larger value of λ will yield a smoother spline, i.e. a smoother estimate of the function that links the observations Y to the design points X.

Question 2 [30 marks]

No code is required for this question.

A data analyst is implementing a Monte Carlo simulation of $M = 1,000$ random samples of realisations of the model

$$Y_i = \theta^* X_i + Z_i, \qquad i = 1, \ldots, n \tag{1}$$

with $n = 100$, $\theta^* = 8$ and a sequence of i.i.d. realizations $Z_i \overset{iid}{\sim} t_d$ with $d = 3$ degrees of freedom, using a single sample $\{X_i\}_{i=1}^n$ from $X \sim \mathcal{U}(1, 2)$ to generate all M Monte Carlo samples, and computes and stores the Monte Carlo least squares estimates of θ^* for analysis. Note the analyst is making sure to not include an intercept in the regression model when fitting it to the simulated data.

(a) Quote the theoretic expected value of Y, i.e. the true value of $E(Y)$, showing your calculation.

(b) Quote the theoretic expected value of $\hat{\theta}$, i.e. the true value of $E(\hat{\theta})$, justifying your answer with a brief statement.

(c) Briefly describe in which way(s) the distribution of Monte Carlo estimates of $\theta^\star$ would differ from the one the analyst is generating, if the additive noise Z was such that $Z \sim \mathcal{N}(0, 1)$, and why (all other settings of the Monte Carlo experiment remaining the same).

(d) Briefly describe in which way(s) the distribution of Monte Carlo estimates of $\theta^\star$ would differ from the one the analyst is generating, if the additive noise Z remained $Z \sim t_d$ with $d = 3$, but using $M = 5,000$ Monte Carlo samples instead of $M = 1,000$, and why (all other settings of the Monte Carlo experiment remaining the same).

(e) The analyst included the instruction

```
shapiro.test(estimates)
```

at the end of the R code, where `estimates` is the vector she used to store the Monte Carlo estimates of $\theta^\star$. Explain what output you would expect from this test, assuming the settings described in (d) above were used for this analysis (i.e. with $M = 5,000$), and why. Provide any information about the test that is relevant to your answer. (No code required.)

(f) Suppose now that the analyst allowed for an intercept in the regression model, when fitting the latter to the Monte Carlo samples simulated correctly from equation (1) above. Briefly describe how this would impact the distribution of estimates of $\theta^\star$ obtained from the original simulation settings described at the beginning of this question, and why.

Solution:

(a) $E(Y) = E(\theta^* X) + E(Z) = \theta^* E(X) + E(Z) = 8 \times 1.5 + 0 = \mathbf{12}$.

(b) The OLS being consistent under this model, $E(\hat{\theta}) = \theta^* = \mathbf{8}$. The MC simulation demonstrates this consistency.

(c)
- The MC mean should remain comparable, since the OLS remains unbiased in this new setting with, again, 0-mean noise.

- The MC estimation variance should decrease, since there would no longer be any outliers in the MC resamples.

(d)
- There would be a larger proportion of outliers in (ie a heavier tail for) the sample of MC estimates because of the heavy-tailed distribution used for Z. However these would be more symmetrically distributed due to the LLN, i.e. the distribution of MC estimates would become more symmetric as M increases.
- Finite-sample variance could increase with M due to more outliers occurring at finite sample horizon, although this increase would be only slight if any, since M is already large enough for the MC distribution to be close to its limit.
- Bias should decrease since $E(\hat{\theta}_M) \to \theta^\star$ as $M \to \infty$.

(e) The p-value should be large enough under $H_0 : \hat{\theta}_M \sim \mathcal{N}(\theta^\star, Var(\hat{\theta}_M))$, since we expect the distribution of `estimates` to be approximately Normally distributed.

(f) Including an intercept in the model would likely increase finite-sample bias slightly, and increase estimation variance noticeably. Even though we'd expect the MC estimates of the intercept to be roughly 0 on average, they would have a non-zero value, and this would "mechanically" affect the accuracy of the estimation of $\theta^\star$.

Question 3 [20 marks]

Note: if you do not manage to answer a question item, provide the R code you would have used, or a comment on the answer you would expect for that question, as relevant.

Load the following library and dataset into your `R` session:

```
library(ISLR)
dat = na.omit(Hitters)
x = dat$Years
y = dat$Salary
```

(a) Fit a nonlinear model of the form

$$y = a + bx + cx^2 + \epsilon$$

to the above data. Quote:

- the coefficient estimates for this model;
- the root mean squared error RMSE corresponding to the model fit.

(b) Fit a cubic smoothing spline to the same data (`x`,`y`), using the default value for the number of degrees of freedom. Quote the RMSE corresponding to this spline model fit.

(c) Fit a second cubic smoothing spline to the same data, using a number of degrees of freedom 4 times higher than the default value used in (b). Quote the RMSE corresponding to this second spline model fit.

(d) Compare the three model RMSEs obtained above, end explain any difference you may observe.

Solution:
R code:

```
# (a.i)
nlo2 = nls(y~a+b*x+c*I(x^2), start=list(a=10,b=1,c=1))
summary(nlo2)
# (a.ii)
(rmse.nlo = sqrt( mean((y.hat-y)^2) ))
# (b)
so = smooth.spline(x,y)
y.so = approx(so$x,so$y,xout=x)$y
(rmse.so = sqrt( mean((y.so-y)^2) ))
# (c)
so2 = smooth.spline(x,y,df=so$df*4)
y.so2 = approx(so2$x,so2$y,xout=x)$y
(rmse.so2 = sqrt( mean((y.so2-y)^2) ))
# (d)
(rmse.so-rmse.nlo)/rmse.nlo
(rmse.so2-rmse.nlo)/rmse.nlo
(rmse.so2-rmse.so)/rmse.so
```

```
# Bonus
plot(x,y)
is = order(x)
lines(x[is],fitted(nlo2)[is],lwd=2,col=2)
lines(so,col=4,lwd=2)
lines(so2,col='navy',lwd=2)
```

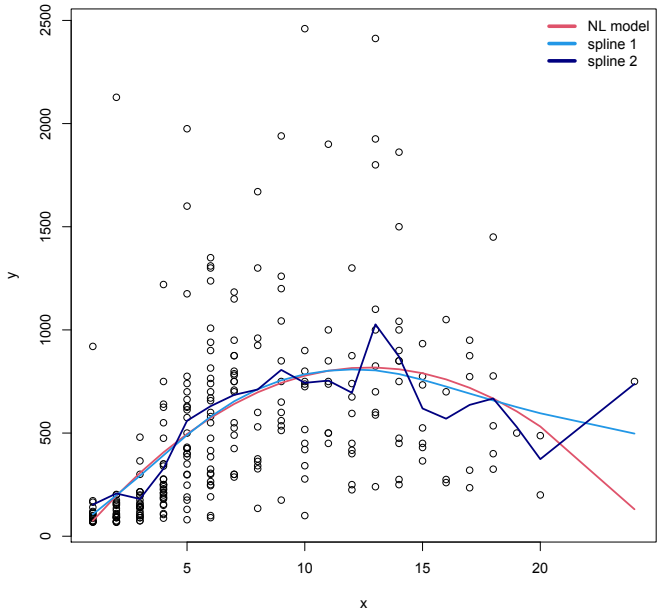
(a) Nonlinear model: a = **-57.442521**, b=**137.626725**, c=**-5.408112**

(b) RMSE of the quadratic model fit: **388.2181**

(c) RMSE of the first smoothing spline: **385.5049**

(d) RMSE of the second smoothing spline: **372.2388**

(e) There is a 0.7% difference between the RMSEs from the parametric model and the first spline, but a 3.4% difference between the RMSEs of the two splines (with a lower RMSE for the second spline). The second spline overfits the data because of its higher number of degrees of freedom, and this yields a reduction in RMSE. One could plot the graph too:



Question 4 [35 marks]

Note: if you do not manage to answer a question item, provide the `R` instruction you would have used, or a comment on the answer you would expect for that question, as relevant.

Load the following library and dataset into your `R` session:

```
library(ISLR)
dat = na.omit(Hitters)
itrain = c(1:200)
dat.train = dat[itrain,]
dat.test = dat[-itrain,]
Salary.test = dat.test$Salary
dat.test$Salary = NULL
```

Set the random seed to 1 (using `set.seed(1)`) before running your analysis. Bootstrap the effect of `HmRun` (the number of home runs in a season) on player salary `Salary`, when measured by univariate linear regression analysis of the training set `dat.train`, use 100 bootstrap resamples of the training set `dat.train`. Record also the corresponding bootstrapped p-values. It is up to you to decide whether an intercept parameter should be included or not.

(a) Run your bootstrap implementation, and:

- Quote the bootstrap mean estimate of the effect of variable `HmRun` on `Salary`.
- Name the quantity estimated in (a).

(b) Quote the 99^{th} percentile of bootstrapped p-values for variable `HmRun`.

(c) Comment on your results for (a) and (b).

(d) Predict the salaries of players in the test set `dat.test` using the means of bootstrap estimates of the univariate regression model parameters. Calculate and quote the root mean squared prediction error.

(e) Comment on your result for (d).

(f) Generate similar predictions for `Salary.test` from the linear regression model obtained by fitting the original dataset `dat.train` (i.e. without resampling). Quote the corresponding root mean squared error.

(g) Comment on your result for (f), in particular what it highlights about the use of bootstrapping to estimate the quantity named in (a).

Solution:
Code:

```
B = 100
set.seed(1)
int = pval = eff = numeric(B)
for(b in 1:B)
ib = sample(1:nrow(dat.train), nrow(dat.train), replace=TRUE)
xb = dat.train[ib,]
```

```
lmb = lm(Salary~HmRun, data=xb)
int[b] = summary(lmb)$coef[1,1]
eff[b] = summary(lmb)$coef[2,1]
pval[b] = summary(lmb)$coef[2,4]

# (a)
mean(eff)
# (c)
quantile(pval,.99)
# (e)
preds = mean(int) + mean(eff)*dat.test$HmRun
sqrt( mean((preds-Salary.test)^2) )
# (f)
mean(Salary.test)
sd(Salary.test)
# (g)
lmo = lm(Salary~HmRun, data=dat.train)
preds.lmo = predict(lmo,dat.test)
sqrt( mean((preds.lmo-Salary.test)^2) )
```

(a)
- Bootstrap mean effect: **18.79508**
- This is an estimate of the expected value of the OLS estimate of the linear parameter in the regression model; i.e. it is an estimate of $E\left[\hat{\theta}_1\right]$ in the regression model $Y = \theta_0 + \theta_1 X + \varepsilon$.

(b) 99-percentile of the p-value of this effect: **0.004532542**

(c) A unit increase in HmRun yields an increase in Salary by a factor of 18.795. This effect is statistically significant.

(d) Test set prediction RMSE from bootstrap model: **364.3579**

(e) The RMSE is very large compared to the mean Salary value (484.7646), and almost equal to its SD (365.7454). Therefore the prediction is not accurate. This variable alone is not sufficient in predicting Salary accurately.

(f) Test set prediction RMSE from original model: **364.9748**

(g) Not surprisingly, the original fit is comparable to the bootstrap estimates. Prediction RMSE is therefore also comparable. This is because bootstrapping provides an unbiased estimate of $E\left[\hat{\theta}_1\right]$.