Selected Topics In Data Science

Bruce Campbell

2020-12-29

Chapter 1

Preface

This is the first installment on my promise to elucidate less popular topics in statistics and machine learning. I wrote this as a way to solidify my understanding of some of the topics that are treated here. Hopefully others will find value here.

Chapter 2

Introduction

This is a living book. It's under development. We are using the **bookdown** package [Xie, 2020] in this book, which was built on top of R Markdown and **knitr** [Xie, 2015].

Chapter 3

On Model Averaging

Recall that we can break down model error into the bias an variance $bias(\hat{Y}) = E[\hat{Y} - E[Y]]$

If we are averaging models $i = 1, \cdots, k$ then

$\text{MSE}\left(\hat{Y}_i\right) = \left\{\text{bias}\left(\hat{Y}_i\right)\right\}^2 + \text{var}\left(\hat{Y}_i\right)$

Chapter 4

Sensitivity Analysis and Shapley Values

Global sensitivity analysis measures the importance of input variables to a function. This is an important task in quantifying the uncertainty in which target variables can be predicted from their inputs. Sobol indices [Saltelli and Sobol', 1995] are a popular approach to this. It turns out that there's a relationship between Sobol indices and Shapley values. We explore this relationship here and demonstrate their effectiveness on some linear and non-linear models.

4.1 Relationship between Sobol indices and Shapley values

Shapley values are based on $f(x) - E[f(x)]$ while Sobol indices decompose output variance into fractions contributed by the inputs. The Sobol index is a global measure of feature importance while Shapley values focus on local explanations although we could combine local Shapley values to achieve a global importance measure. Sobol indices are based on expectations and can be used for features not included in the model / function of interest. In this way we could query for important features correlated with those that the model does use.

4.2 CRAN sensitivity package

```r
library(ggplot2)
library(pander)
if(!require(sensitivity)){
    install.packages("sensitivity")
    library(sensitivity)
}
```

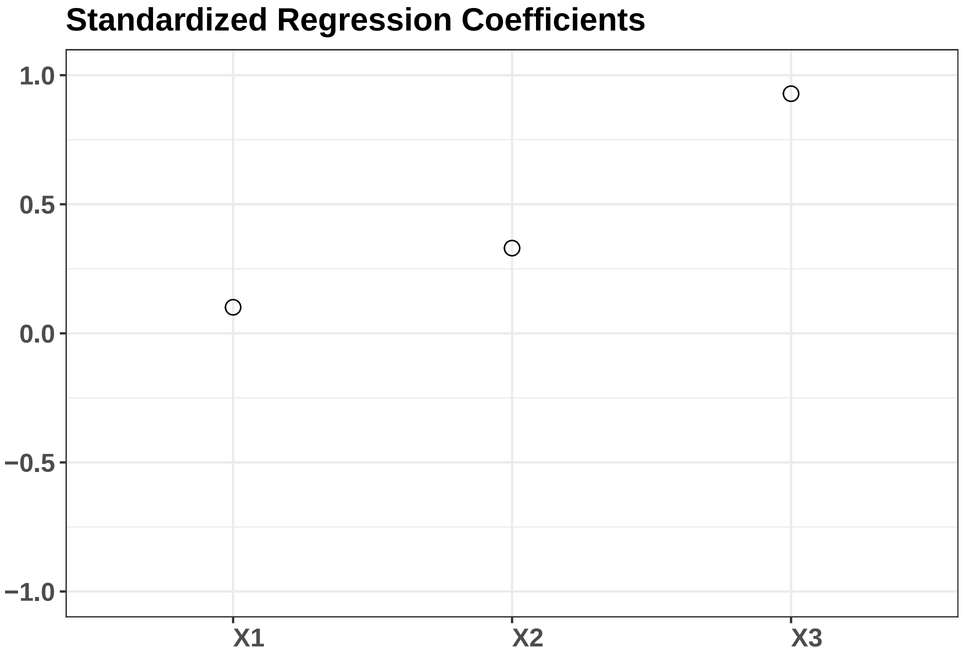
Standardized Regression Coefficients (SRC), or the Standardized Rank Regression Coefficients (SRRC), which are sensitivity indices based on linear or monotonic assumptions in the case of independent factors.

```r
n <- 100
X <- data.frame(X1 = runif(n, 0.5, 1.5),
                X2 = runif(n, 1.5, 4.5),
                X3 = runif(n, 4.5, 13.5))

# linear model : Y = X1 + X2 + X3

y <- with(X, X1 + X2 + X3)



Z <- src(X, y, rank = FALSE, logistic = FALSE, nboot = 0, conf = 0.95)

pander(Z$SRC,caption = "Standardized Regression Coefficients ")
```

Table 4.1: Standardized Regression Coefficients

	original
X1	0.101
X2	0.3301
X3	0.9283

```r
ggplot(Z, ylim = c(-1,1))+ggtitle("Standardized Regression Coefficients")
```

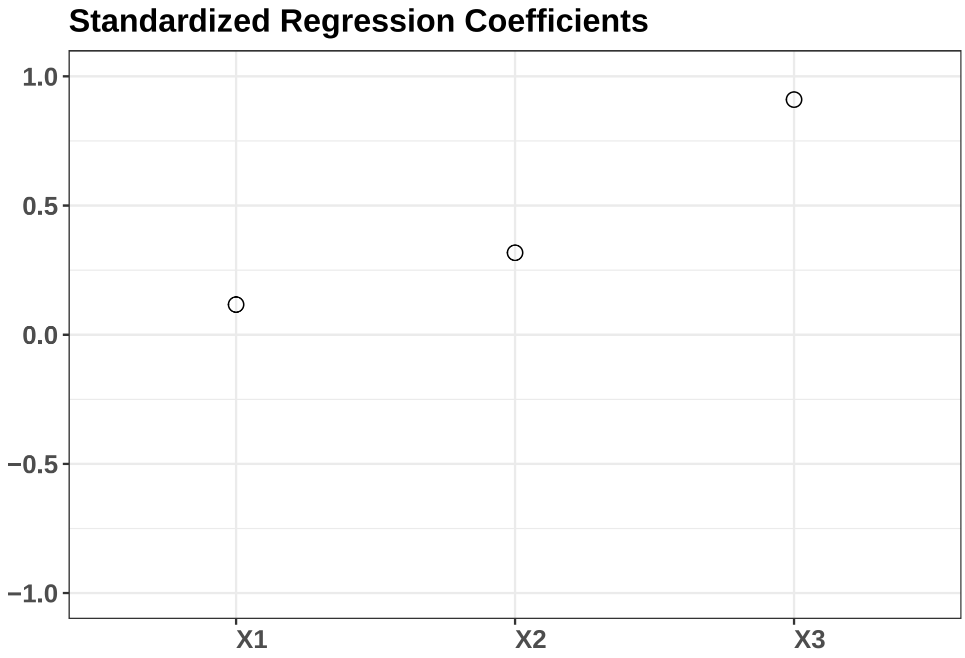


```r
y <- with(X, X1 + X2 + X3)
y <- y + rnorm(nrow(X),0,1/2)
df<- data.frame(cbind(X,y))

Z <- src(X, y, rank = FALSE, logistic = FALSE, nboot = 0, conf = 0.95)

pander(Z$SRC,caption = "Standardized Regression Coefficients ")
```

Table 4.2: Standardized Regression Coefficients

	original
X1	0.1165
X2	0.3171
X3	0.9099

```
ggplot(Z, ylim = c(-1,1))+ggtitle("Standardized Regression Coefficients")
```



```
#lm.fit = lm(y ~ X1+X2+X3,data = df)
#summary(lm.fit)
#attach(df)
#plot(y, X1+X2+X3)
```

We see how the importance of X3 is ranked above X2 and likewise X2 is more important than X1. This is by design of the simulated data set. The standardized regression coefficients (beta coefficients) are calculated from that has been standardized, let's normalize and calculate the regression to see if indeed that is the case.

```
dfs<- data.frame(scale(df,center = TRUE,scale = TRUE))
lm.fit = lm(y ~ X1+X2+X3,data = dfs)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = y ~ X1 + X2 + X3, data = dfs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62741 -0.10185  0.01695  0.14604  0.38982
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.264e-16  1.936e-02   0.000        1
## X1          1.165e-01  1.986e-02   5.863 6.42e-08 ***
## X2          3.171e-01  1.985e-02  15.976  < 2e-16 ***
## X3          9.099e-01  1.947e-02  46.735  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1936 on 96 degrees of freedom
## Multiple R-squared:  0.9637, Adjusted R-squared:  0.9625
## F-statistic: 848.7 on 3 and 96 DF,  p-value: < 2.2e-16
```

We see that the values are very close.

4.3 Partial Correlation Coefficients

```
x <- pcc(X, y, nboot = 100)
print(x)
```

```
##
```

```
## Call:
## pcc(X = X, y = y, nboot = 100)
##
## Partial Correlation Coefficients (PCC):
##     original          bias  std. error min. c.i. max. c.i.
## X1 0.5134715  5.023519e-04 0.066248670 0.4157600 0.6870635
## X2 0.8524548 -3.249625e-06 0.021846339 0.8139257 0.9052334
## X3 0.9787221 -3.341981e-05 0.003358137 0.9730419 0.9856168
```

4.4 Sobol indices for deterministic function and for model

```
y.fun <- function(X) {

  X1<- X[,1]
  X2<- X[,2]
  X3<- X[,3]

  X1+X2+X3
}

yhat.fun<-function(X,lm)
{
  X1<- X[,1]
  X2<- X[,2]
  X3<- X[,3]

  yhat <- predict(lm.fit,data.frame(X1=X1,X2=X2,X3=X3))
  return(yhat)
}

nboot = 100
```

```
x <- sobol(model = y.fun, X[1:50,], X[51:100,], order = 2, nboot = nboot)
S.sobol <- x$S
pander(S.sobol)
```

	original	bias	std. error	min. c.i.	max. c.i.
X1	0.01409	-0.03702	0.8124	-1.581	2.173
X2	-0.1993	-0.05273	0.818	-1.678	1.794
X3	1.63	0.05383	0.3927	0.8466	2.476
X1*X2	0.014	0.02349	0.7954	-2.086	1.583
X1*X3	0.014	0.02349	0.7954	-2.086	1.583
X2*X3	0.014	0.02349	0.7954	-2.086	1.583

```
#yhat.fun(data.frame(X1=1,X2=2,X3=3),lm.fit)

x <- sobol(model = yhat.fun,X[1:50,], X[51:100,], order = 2, nboot = nboot)
S.sobol <- x$S
pander(S.sobol)
```

	original	bias	std. error	min. c.i.	max. c.i.
X1	-0.2086	-0.03056	0.7721	-1.751	1.523
X2	-0.2831	-0.06508	0.7892	-1.766	1.478
X3	1.383	0.04754	0.1276	1.005	1.557
X1*X2	0.2113	0.03101	0.7722	-1.516	1.746
X1*X3	0.2113	0.03101	0.7722	-1.516	1.746
X2*X3	0.2113	0.03101	0.7722	-1.516	1.746

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.

Bibliography

Andrea Saltelli and I.M. Sobol'. Sensitivity analysis for nonlinear mathematical models: Numerical experience. *Matematicheskoe Modelirovanie*, 7, 01 1995.

Yihui Xie. *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition, 2015. URL http://yihui.org/knitr/. ISBN 978-1498716963.

Yihui Xie. *bookdown: Authoring Books and Technical Documents with R Markdown*, 2020. URL https://github.com/rstudio/bookdown. R package version 0.21.