L1Workshop1

Ye

6/28/2017

Linear regression with number of independent predictors from 2 to 500.

$Y_{i,j} = \beta_0 + \beta_1 X_{i,1} + ... + \beta_j X_{i,j} + \epsilon_i; i = 1, ..., 500; j = 2, ..., 500.$

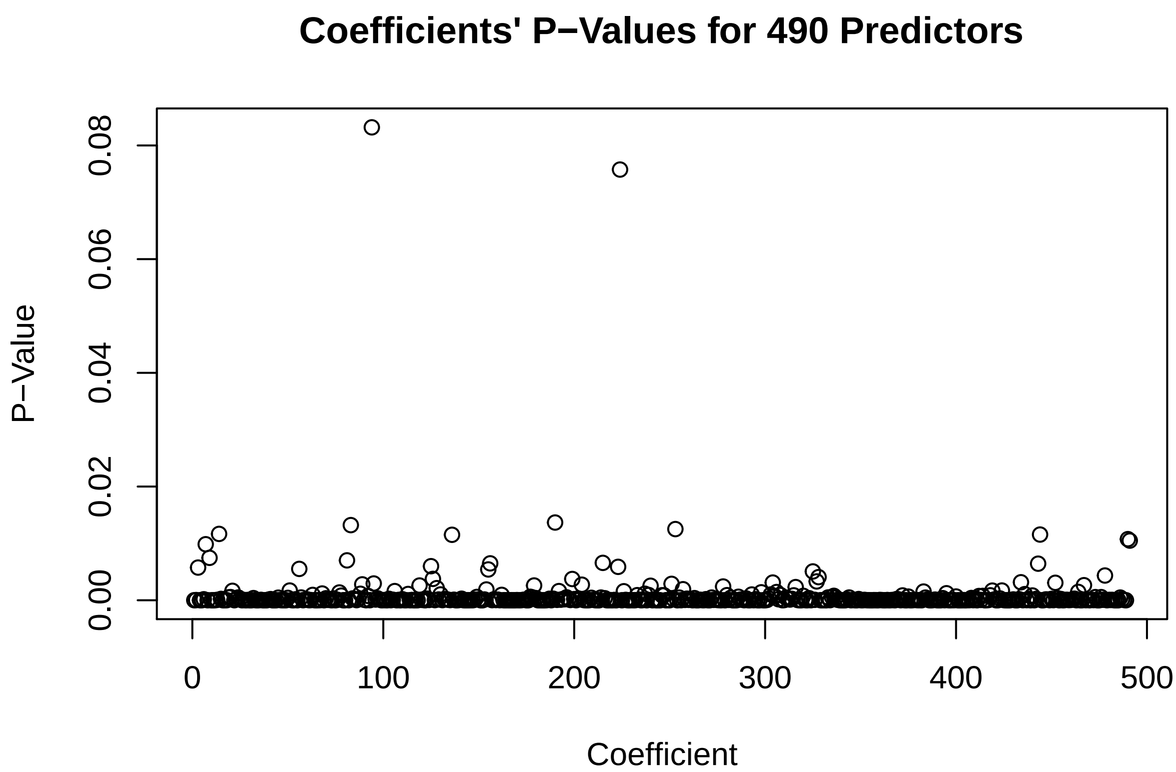
```
set.seed(8394756)
Epsilon = rnorm(500, 0, 1)
X = rnorm(500*500, 0, 2)
dim(X) = c(500, 500)
colnames(X) = paste0("X", 1:500)
slopesSet = runif(500, 1, 3)
Y = sapply(2:500, function(z) 1 + X[, 1:z] %*% slopesSet[1:z] + Epsilon)
```

Analysis of accuracy of inference as a function of number of the predictors

```
completeModelDataFrame = data.frame(Y=Y[,490], X[, 1:491])
m2 = lm(Y[,1]~X[,1:2])
m490 = lm(Y~., data=completeModelDataFrame)
summary(m2)
```

```
##
## Call:
## lm(formula = Y[, 1] ~ X[, 1:2])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.70335 -0.62160  0.04297  0.63964  2.76616
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.00267    0.04444   22.56   <2e-16 ***
## X[, 1:2]X1   2.68339    0.02117  126.75   <2e-16 ***
## X[, 1:2]X2   2.29734    0.02321   98.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9935 on 497 degrees of freedom
## Multiple R-squared:  0.9812, Adjusted R-squared:  0.9811
## F-statistic: 1.294e+04 on 2 and 497 DF,  p-value: < 2.2e-16
```

```
plot(coefficients(summary(m490))[-1,4],
     main="Coefficients' P-Values for 490 Predictors",
     xlab="Coefficient",
     ylab="P-Value")
```



Both summaries show pretty strong significanceof all predictors. #Check 95% confidence intervals for the first predictor $X_{i,1}$ estimated for both models.

```
confint(m2)[2,]
```

```
##    2.5 %   97.5 %
## 2.641792 2.724980
```

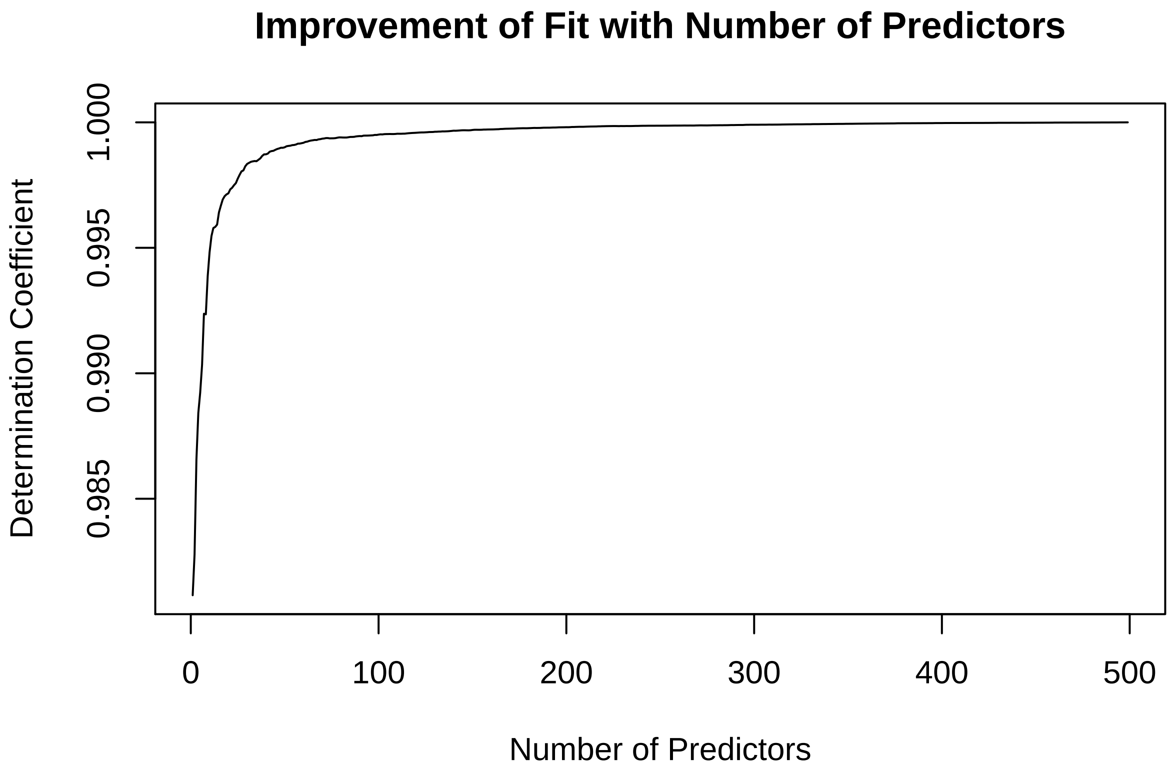
```
confint(m490)[2,]
```

```
##    2.5 %   97.5 %
## 1.927847 3.004426
```

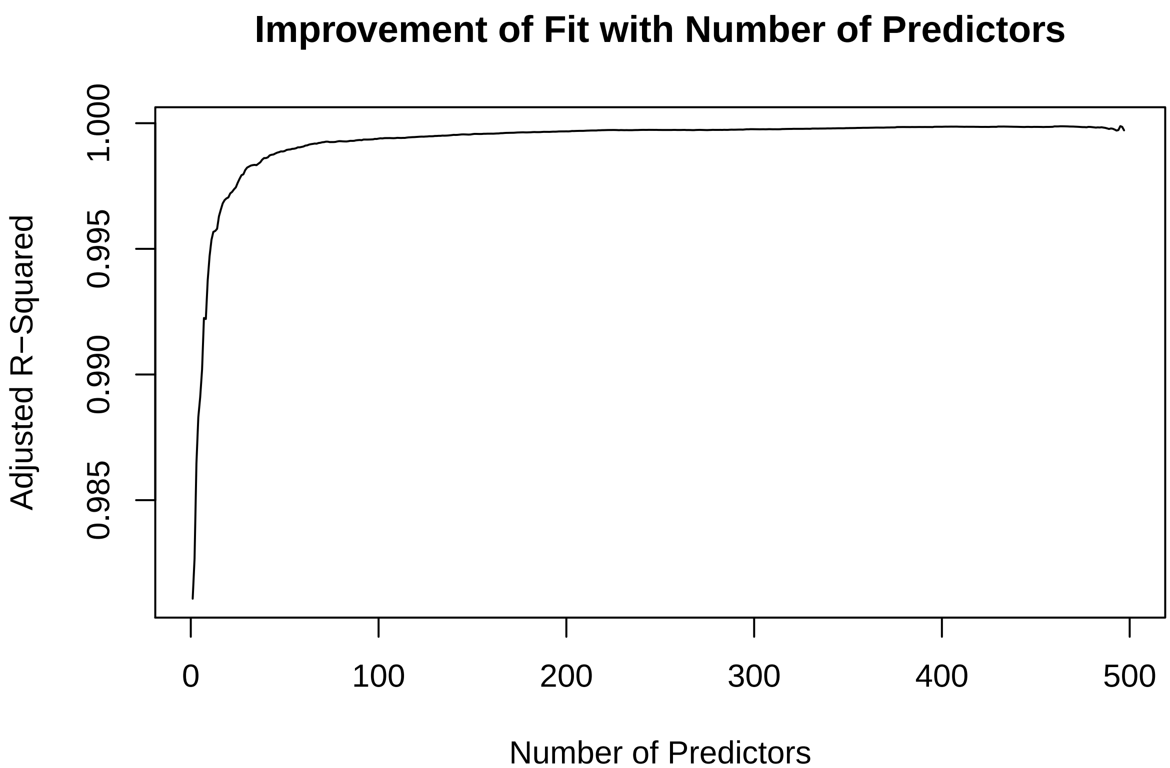
Explain?

Coefficients of determination (R^2) and adjusted $R^2 = 1 - \frac{SSE/(n-k)}{SST/n-1}$

```
rSquared<-sapply(2:500,function(z) summary(lm(Y~.,data=data.frame(Y=Y[,z-1],X[,1:z])))$r.squared)
plot(rSquared,type="l",
     main="Improvement of Fit with Number of Predictors",
     xlab="Number of Predictors",
     ylab="Determination Coefficient")
```



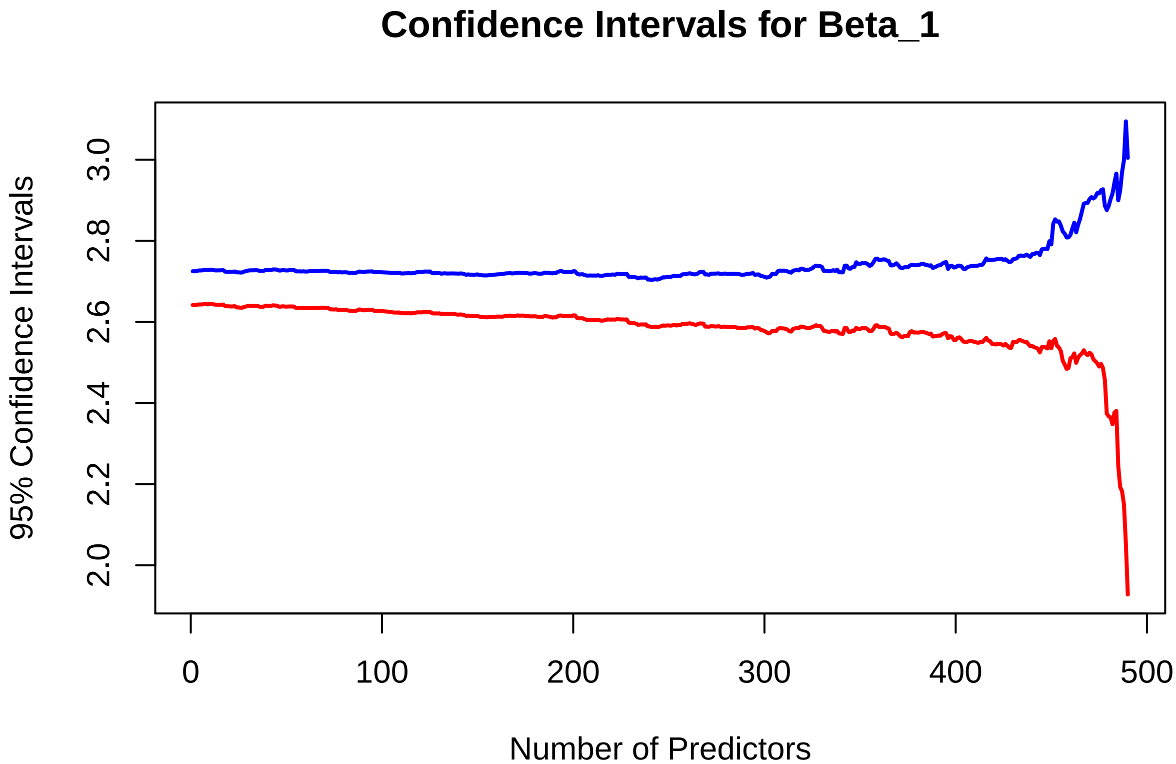

```r
adjustedRSquared<-sapply(2:500,function(z) summary(lm(Y~.,data=data.frame(Y=Y[,z-1],X[,1:z])))$adj.r.squ
plot(adjustedRSquared,type="l",
     main="Improvement of Fit with Number of Predictors",
     xlab="Number of Predictors",
     ylab="Adjusted R-Squared")
```



#

Plot the confidence interval of $X_{i,1}$ for all nested models

```
leftConfInt<-suppressWarnings(sapply(2:500,function(z) confint(lm(Y~.,data=data.frame(Y=Y[,z-1],X[,1:z]
rightConfInt<-suppressWarnings(sapply(2:500,function(z) confint(lm(Y~.,data=data.frame(Y=Y[,z-1],X[,1:z
matplot(1:490,cbind(leftConfInt[1:490],rightConfInt[1:490]),type="l",lty=1,
        lwd=2,col=c("red","blue"),main="Confidence Intervals for Beta_1",
        xlab="Number of Predictors",ylab="95% Confidence Intervals")
```



#

Conclusions: 1. As number of predictors grows the quality of fit expressed as R^2 or adjusted R^2 continuously improves. 2. But inference for a fixed predictor becomes less and less accurate, which is shown by the widening confidence interval. 3. This means that if there is, for example, one significant predictor $X_{i,1}$, by increasing the total number of predictors (even though they all or many of them may be significant) we can damage accuracy of estimation of the slope for $X_{i,1}$. 4. This example shows one problem that DM has to face, which is not emphasized in traditional courses on statistical analysis where only low numbers of predictors are considered.

Selecting predictors for regresssion (drop1() or step())

```
m10<-lm(Y~.,data=data.frame(Y=Y[,9],X[,1:10]))
(drop1.m10<-drop1(m10))
```

```
## Single term deletions
##
## Model:
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
##        Df Sum of Sq     RSS     AIC
## <none>                485.5    7.30
## X1      1   15571.2 16056.7 1754.64
## X2      1    9532.1 10017.6 1518.75
```

```
## X3       1    2484.5  2970.1  910.87
## X4       1    6990.5  7476.0 1372.43
## X5       1    8850.5  9336.0 1483.51
## X6       1    2005.7  2491.3  822.97
## X7       1    5787.9  6273.4 1284.73
## X8       1   10603.3 11088.8 1569.54
## X9       1    1898.7  2384.3  801.02
## X10      1   15388.7 15874.2 1748.92
```

```
bestToDrop<-drop1.m10[which.min(drop1.m10$AIC),]
(step.m10<-step(m10,direction="both"))
```

```
## Start:  AIC=7.3
## Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10
##
##        Df Sum of Sq     RSS     AIC
## <none>                485.5    7.30
## - X9    1    1898.7  2384.3  801.02
## - X6    1    2005.7  2491.3  822.97
## - X3    1    2484.5  2970.1  910.87
## - X7    1    5787.9  6273.4 1284.73
## - X4    1    6990.5  7476.0 1372.43
## - X5    1    8850.5  9336.0 1483.51
## - X2    1    9532.1 10017.6 1518.75
## - X8    1   10603.3 11088.8 1569.54
## - X10   1   15388.7 15874.2 1748.92
## - X1    1   15571.2 16056.7 1754.64
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 +
##     X10, data = data.frame(Y = Y[, 9], X[, 1:10]))
##
## Coefficients:
## (Intercept)           X1           X2           X3           X4
##       1.001        2.685        2.296        1.089        1.859
##          X5           X6           X7           X8           X9
##       2.074        1.031        1.656        2.389        1.004
##         X10
##       2.831
```