Final Quiz Solution

Modern Data Mining/Linda

April 29, 2021

Instruction: This is an open book, 30-minute quiz.

Statistics Concepts

1. For any predictive model built from a dataset, what is more appropriate metric to consider?

(A) Training Error

(B) Testing Error

Answer: (B) Since we are building a predictive model, testing error will be the key to be minimized.

2. For linear regression with more than 10 covariates (or features), which of the following is TRUE?

(A) LASSO (with `lambda.1se`) and minimizing C_p will choose the same set of variables.

(B) LASSO (with `lambda.min`) and minimizing C_p will choose the same set of variables.

(C) The model with the smallest C_p is expected to perform well on testing data.

(D) None of the above is TRUE.

Answer: (C). While C_p estimate the testing error but the metric is different from that of LASSO target function. We expect the best model selected are similar but there is no guarantee the same.

3. In case of logistic regression, deviance (negative log-likelihood) never increases with an addition of one more covariate.

(A) TRUE

(B) FALSE

Answer: (A) deviance being - loglikelihood which is similar to that of SSE in regression. The more variables we add the smaller deviance will become.

Part 1. EDA / PCA

We will use the IQ data from the lecture and homework. The response variable of interest is `Income2005`. We take the log transformation `log(Income2005)` and create a new variable `Income2005_trans`.

```
data_IQ <- read.csv("IQ.Full.csv")  # We will use IQ data set
data_IQ$Race <- as.factor(data_IQ$Race)
data_IQ$Income2005_trans <- log(data_IQ$Income2005)
```

4. We first perform PCA. We only focus on the following variables: MotherEd, FatherEd, Educ, Science, Word, and Auto.

```
data_pca <- data_IQ %>% dplyr::select(MotherEd, FatherEd, Educ, Science, Word,
    Auto)
pc <- prcomp(data_pca, scale = TRUE, center = TRUE)
pc$rotation
```

```
##            PC1    PC2    PC3    PC4    PC5      PC6
## MotherEd 0.383  0.462 -0.380  0.402 -0.577  0.00789
## FatherEd 0.386  0.458 -0.368 -0.396  0.590  0.01900
## Educ     0.377  0.268  0.723  0.422  0.283 -0.06927
## Science  0.476 -0.329  0.109 -0.184 -0.136  0.77527
## Word     0.470 -0.208  0.197 -0.495 -0.337 -0.58138
## Auto     0.338 -0.594 -0.382  0.473  0.327 -0.23607
```

```
pc.s = summary(pc)
round(pc.s$importance, 3)
```

```
##                          PC1   PC2   PC3   PC4   PC5   PC6
## Standard deviation     1.80 1.047 0.826 0.638 0.612 0.443
## Proportion of Variance 0.54 0.183 0.114 0.068 0.062 0.033
## Cumulative Proportion  0.54 0.723 0.837 0.905 0.967 1.000
```

```
data_final <- data.frame(Income2005 = data_IQ$Income2005, data_pca, PC1 = pc$x[,
    1], PC2 = pc$x[, 2])

### correlation between LogIncome and all other variable
t(cor(log(data_IQ$Income2005), data_final))
```

```
##               [,1]
## Income2005  0.7515
## MotherEd    0.1648
## FatherEd    0.1905
## Educ        0.2881
## Science     0.2660
## Word        0.2319
## Auto        0.3173
## PC1         0.3267
## PC2        -0.0796
```

Based on outputs from R-chunks above, choose the CORRECT statements.

(A) PC1 score is proportional to the sum of weighted six variables.

(B) PC1 and PC2 together take over 70% of the total variability.

(C) Based on the correlation table of `Income2005_trans` vs. all the variables listed there, the best **single** sensible variable linear model to predict `Income2005_trans` will be Income2005. (with highest R square.)

(D) Based on the correlation table of `Income2005_trans` vs. all the variables listed there, the best **single** sensible variable linear model to predict `Income2005_trans` will be PC1. (with highest R square.)

Answer: (A), (B) and (D). (c) is conceptually WRONG!!! One CAN't use response to predict future response!!!!

Part 2. Regression

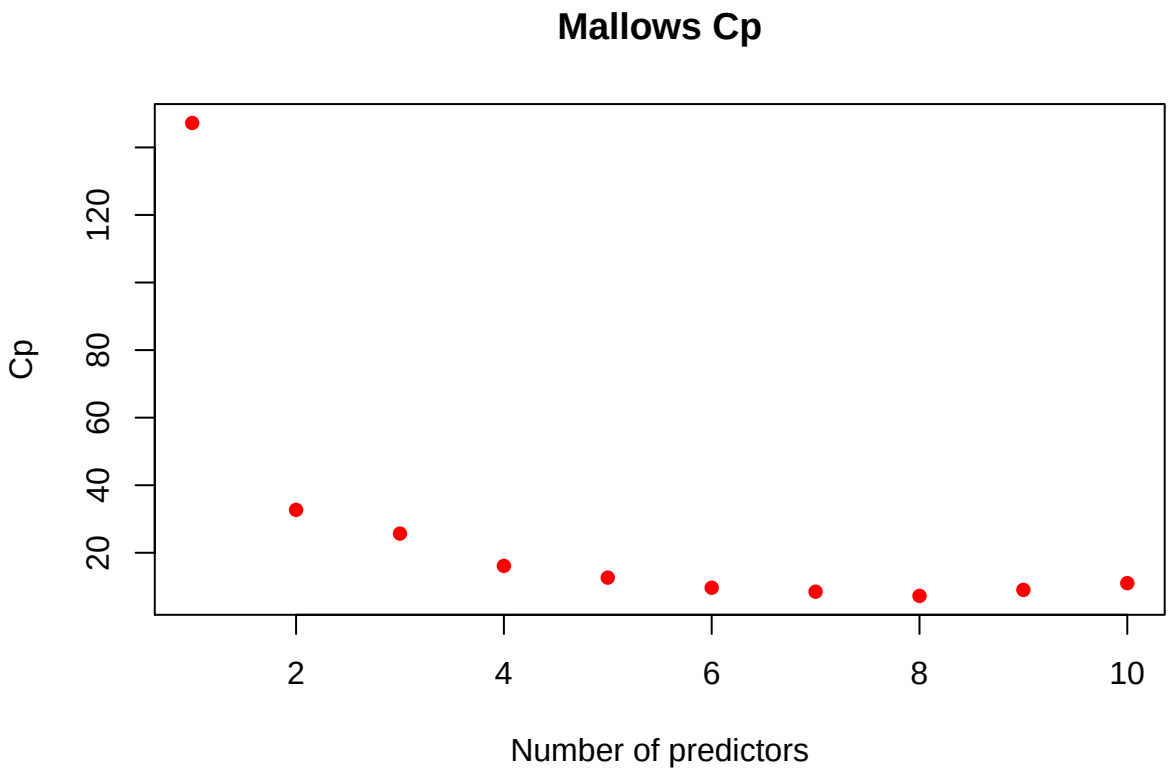
From here, we will focus on the transformed income **Income2005_trans**. We are interested in the best features that predict the **Income2005_trans**, out of Science, Arith (Arithmetic reasoning), Word (Word knowledge), Parag (Paragraph comprehension), Numer (Numerical operation), Coding (Coding speed), Auto (Automative and Shop information), Math (Math knowledge), Mechanic (Mechanic Comprehension) and Elec (Electronic information). We run `regsubsets` in the following chunk.

```
fit.exh <- regsubsets(Income2005_trans ~ Science + Arith + Word + Parag + Numer +
    Coding + Auto + Math + Mechanic + Elec, data = data_IQ, nvmax = 10, method = "exhaustive")
fit.sum <- summary(fit.exh)
fit.sum$which[, -1]
```

```
   Science Arith  Word Parag Numer Coding Auto  Math Mechanic  Elec
1    FALSE FALSE FALSE FALSE FALSE  FALSE TRUE FALSE    FALSE FALSE
2    FALSE FALSE FALSE FALSE FALSE  FALSE TRUE  TRUE    FALSE FALSE
3    FALSE FALSE FALSE FALSE  TRUE  FALSE TRUE  TRUE    FALSE FALSE
4    FALSE FALSE FALSE FALSE  TRUE   TRUE TRUE  TRUE    FALSE FALSE
5    FALSE FALSE  TRUE FALSE  TRUE   TRUE TRUE  TRUE    FALSE FALSE
6    FALSE  TRUE  TRUE FALSE  TRUE   TRUE TRUE  TRUE    FALSE FALSE
7    FALSE  TRUE  TRUE FALSE  TRUE   TRUE TRUE  TRUE    FALSE  TRUE
8     TRUE  TRUE  TRUE FALSE  TRUE   TRUE TRUE  TRUE    FALSE  TRUE
9     TRUE  TRUE  TRUE  TRUE  TRUE   TRUE TRUE  TRUE    FALSE  TRUE
10    TRUE  TRUE  TRUE  TRUE  TRUE   TRUE TRUE  TRUE     TRUE  TRUE
```

We now use C_p to choose among models with different number of predictors.

```
plot(fit.sum$cp, xlab = "Number of predictors", ylab = "Cp", col = "red", type = "p",
    pch = 16, main = "Mallows Cp")
```



5. Based on two outputs above, choose the CORRECT statement.

(A) C_p estimates the testing error for each model.
(B) If we want to choose a model with 4 variables based on the C_p plot, the selected variables will be `Numer`, `Coding`, `Auto` and `Math`.
(C) If we want to choose a model with 4 variables based on the C_p plot, we need more information to determine which 4 variables are chosen.
(D) If we want to choose a model with small testing error, R squared would not be an appropriate criterion.

Answer: (A), (B) and (D). Given a model with a fixed number of predictors, C_p estimate the testing errors for the best model (smallest RSS) of the same size. That best model for each size is reported from regsubsets(). So if we decided to use a 4 variables based on the C_p it has to be the one reported already above.

Part 3. Logistic Regression/Classification

We now switch gear to explore the relationship between the reported Family Income at the year 1978, i.e., the variable **FamilyIncome78**, and the demographics. In particular, we know the mean family income in 1978 in US is \$15060. We create a new variable `Above_Mean_Income`, which is equal to 1 if **FamilyIncome78** $\geq$ 15060 and 0 otherwise. We fit a logistic regression as shown below:

```
data_binary <- data_IQ
data_binary$Above_Mean_Income <- as.numeric(data_IQ$FamilyIncome78 >= 15060)
fit_logistic <- glm(Above_Mean_Income ~ MotherEd + FatherEd + Gender + Race,
    data_binary, family = binomial())
print(xtable(fit_logistic), type = output_format)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.7455	0.2551	-10.76	0.0000
MotherEd	0.0972	0.0219	4.44	0.0000
FatherEd	0.1162	0.0158	7.34	0.0000
Gendermale	0.0834	0.0855	0.98	0.3292
Race2	-0.1205	0.2311	-0.52	0.6021
Race3	0.7966	0.1924	4.14	0.0000

6. Based on the summary table, choose the CORRECT statements.

(A) `Race` is a categorical variable with 3 levels. When fitting a glm regression, we only see two levels from the summary table because no subject has Race1 in the data.

(B) The estimate of the coefficients is obtained by minimizing the negative log-likelihood function.

(C) The chance of being `Above_Mean_Income` increases when Mother's education is higher holding all other features the same.

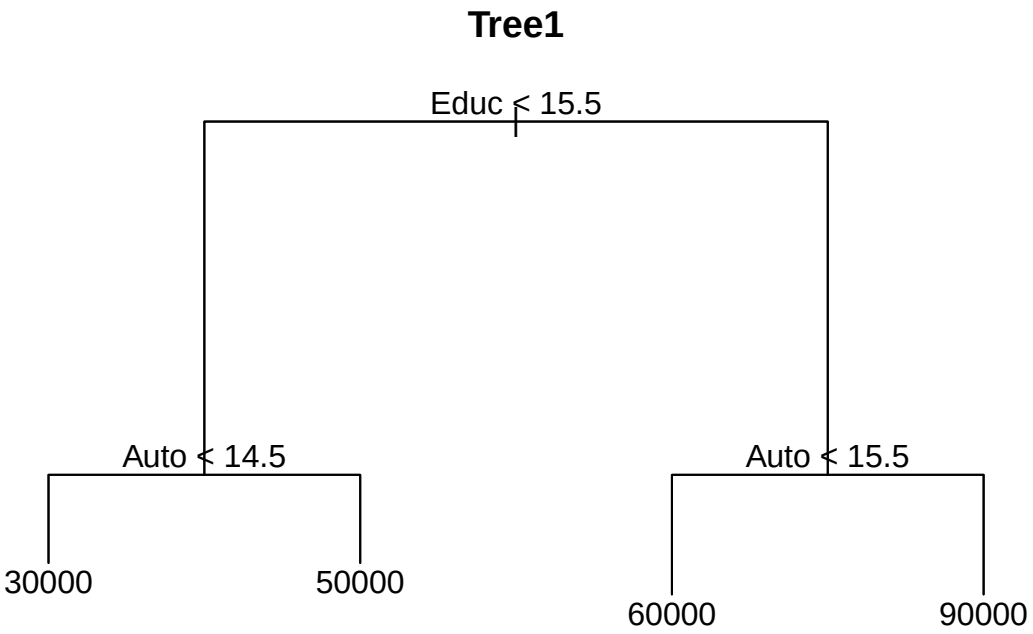
(D) Male is more likely to get higher income than female holding all other features the same.

Answer: (B), (C) and (D). For any categorical variable with K levels, the base (intercept) represents the K-th level.

Part 4. Trees / Bagging / Random Forest

We now explore some tree-based methods for predicting the income in the year of 2005. We include some variables of interest and build a single tree.

```
data_tree <- data_IQ %>% dplyr::select(MotherEd, FatherEd, FamilyIncome78, Educ,
    Science, Arith, Word, Parag, Numer, Coding, Auto, Math, Mechanic, Elec,
    Income2005)
singletree1 <- tree(Income2005 ~ ., data_tree)
# plot(singletree,type='uniform')
plot(singletree1)
text(singletree1)
title(main = "Tree1")
```



7. Based on the tree1, which statements are CORRECT?

(A) Income2005 is predicted to be higher if `Auto` score is higher.

(B) Income2005 is predicted to be higher if `Auto` score is higher while holding all other variables to be the same.

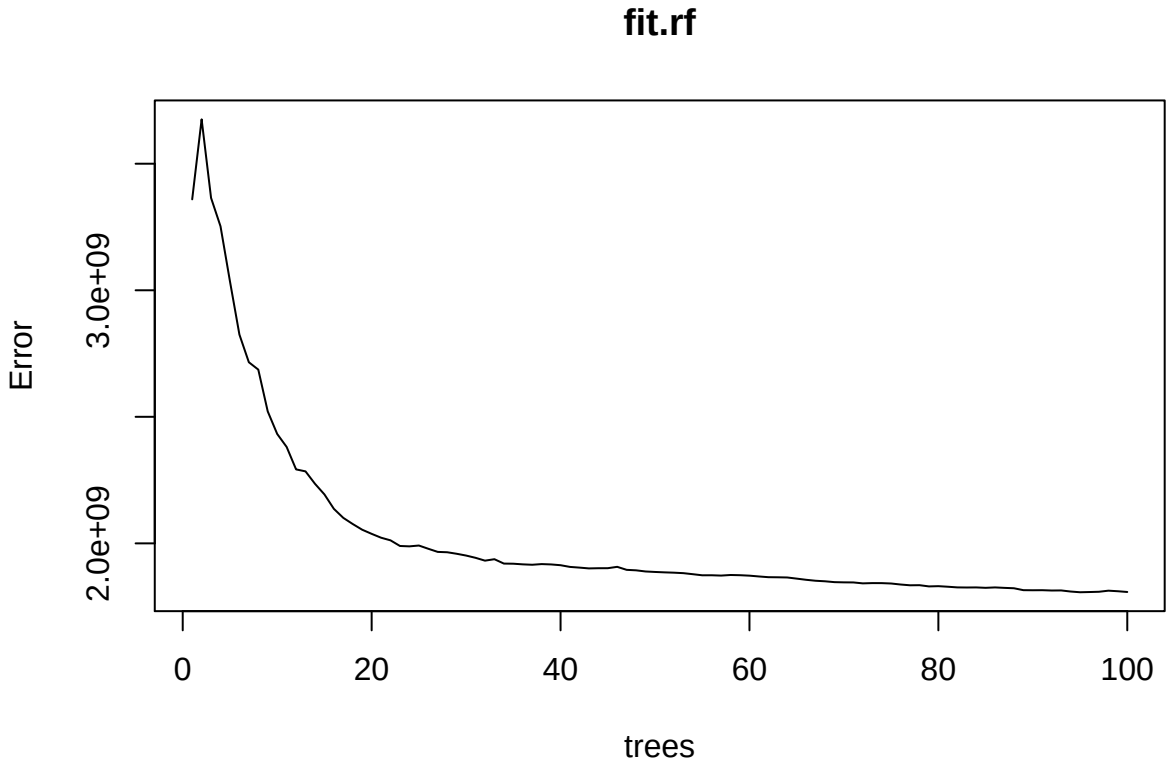
(C) The above decision tree is not a linear model in `Auto` and `Educ`.

(D) There are only 4 predicted values based on this model.

Answer: (B), (C) and (D). (A) is not true because the outcome depends on Educ.... But holding Educ to be the same then we see that the higher Auto is, the larger predictive values become which can be observed from the tree.

We next build a random forest using the same data.

```
fit.rf <- randomForest(Income2005 ~ ., data_tree, mtry = 3, ntree = 100)
plot(fit.rf)
```



8. For the `fit.rf` model above, choose the best description that best describes how this model was built.

(A) Randomly sampling subjects with replacement and bagging 100 trees with the 3 best features for each tree.

(B) Randomly sampling 3 features and bagging 100 bootstrap trees on all subjects

(C) Randomly sampling subjects with replacement and bagging 100 trees and 3 features are randomly sampled at each split for each tree.

Answer: (C) describes RandomForest

9. Which statement is CORRECT

(A) The testing error fit.rf with 100 trees is smaller than that of fit.rf with 60 trees.

(B) In general the training error of fit.rf with 100 trees is always smaller than that of fit.rf with 60 trees.

Answer: (A) The plot shows testing errors of RF as a function of size of the bag, namely the number of trees being averaged. Because of the algorithm of RF, there is no guarantee about training errors of any sort.

Part 5: Neural Network

10. We focus on the Yelp review data with Neural Network.

The data structure is as follows: word frequencies of 1032 words for each review is extracted as the predictors, the response variable y is a binary of review rating, 1 indicates good and 0 bad.

The following R-chunk specifies a neural network to model P(y=1|a review) and P(y=0|a review)

```
model <- keras_model_sequential() %>% layer_dense(units = 16, activation = "relu",
    input_shape = c(1032)) %>% layer_dense(units = 16, activation = "relu") %>%
    layer_dense(units = 8, activation = "relu") %>% layer_dense(units = 2, activation = "softmax")   # o
```

Which statements are CORRECT?

(A) The model is different from a logistic regression model since log odds of $y = 1$ is not a linear function of the predictors.

(B) $P(y = 1|x_1, \ldots, x_{1032})$ can be calculated given the values of the last layer before the outputs.

(C) There are three layers and each layer has 16, 16 and 8 neurons respectively. Each neuron is created by taking the linear combination of all the neurons from the previous layer then apply relu function.

(D) The model or the architecture is a two layer model with 16+16+8=40 and 2 neurons in each layer.

Answer: (A), (B), (C).

(A): The output layer is a function for input but with several layer of nonlinear transformation vie a nonlinear function relu. So the logodds is not longer a linear function of the input any more.

(B) By construction we see that the output layer only depends on the previous layer.

(C) That is the correct terminology in Neural Network!

END