Test 2 Practice

DSO 530: Applied Modern Statistical Learning Methods

2020

You have 120 minutes to do the problems. For multiple choice questions (1-25), make sure to read the questions very carefully and circle the best answer. If you circle multiple answers for a question, you will receive zero for that question. For short answer questions (26-30), write concisely and clearly. This test is open notes, but no electronic equipment other than a calculator is allowed. Questions 26-30 are worth 2 points each, and the rest are worth 1 point each. The total points are 35.

This is a real test from a previous year. The R coding questions are not relevant for the current year. Also, ignore question 26.

Name (Last, First):

USC Student ID number:

Sign on the line below to pledge that you did not give or receive assistance on the questions for this exam from anyone else.

---

part a) multiple choices

1.

```
set.seed(1); x = rnorm(100); set.seed(1); y = -0.5 * rnorm(100)
```

What is the correlation between x and y?

- A) 0.5
- B) -0.5
- C) 1
- D) -1
- E) It cannot be determined given the above information.

Answer: D)

2. Which of the following is/are true about correlation r?

i) It measures linear relationship between two numerical variables

ii) It takes values between -1 and 1

iii) If $r = 0.9$, the strength of the linear relationship is the same as that when $r = -0.9$.

- A) i), ii), iii)
- B) ii) and iii)
- C) i) ii)

- D) ii)
- E) iii)

Answer: A)

3. Which of the following is/are correct?

i) LASSO is always a better method than ridge regression.

ii) LASSO tends to select sparser models compared to ridge regression.

iii) LASSO and ridge regression cannot be applied to the situations where the number of features is bigger than the sample size.

- A) i), ii), iii)
- B) ii) and iii)
- C) i) ii)
- D) ii)
- E) iii)

Answer: D)

4. Which of the following is/are TRUE about confidence intervals?

i) For the confidence intervals, the wider the better.

ii) Other things being equal, 99% CI is wider than the 95% CI

iii) Confidence intervals and prediction intervals are both interval estimates for parameters.

- A) i)
- B) ii)
- C) iii)
- D) i), ii)
- E) ii), iii)
- F) i), ii), iii)

Asnwer: B)

5. Which of the following are/is true?

i) In general, we prefer C_p statistics small and adjusted R squares large.

ii) Backward stepwise regression cannot be used when the number of features is larger than the sample size.

iii) Best subset selection is a computationally intensive method when the number of features is large.

- A) i)
- B) ii)
- C) iii)
- D) i), ii)
- E) i), iii)
- F) ii), iii)
- G) i), ii), iii)

Answer: G).

6.

```
library(ISLR); attach(Smarket); summary(Smarket);
```

```
##       Year           Lag1                Lag2
##  Min.   :2001   Min.   :-4.922000   Min.   :-4.922000
##  1st Qu.:2002   1st Qu.:-0.639500   1st Qu.:-0.639500
##  Median :2003   Median : 0.039000   Median : 0.039000
##  Mean   :2003   Mean   : 0.003834   Mean   : 0.003919
##  3rd Qu.:2004   3rd Qu.: 0.596750   3rd Qu.: 0.596750
##  Max.   :2005   Max.   : 5.733000   Max.   : 5.733000
##       Lag3                Lag4                Lag5
##  Min.   :-4.922000   Min.   :-4.922000   Min.   :-4.92200
##  1st Qu.:-0.640000   1st Qu.:-0.640000   1st Qu.:-0.64000
##  Median : 0.038500   Median : 0.038500   Median : 0.03850
##  Mean   : 0.001716   Mean   : 0.001636   Mean   : 0.00561
##  3rd Qu.: 0.596750   3rd Qu.: 0.596750   3rd Qu.: 0.59700
##  Max.   : 5.733000   Max.   : 5.733000   Max.   : 5.73300
##      Volume           Today           Direction
##  Min.   :0.3561   Min.   :-4.922000   Down:602
##  1st Qu.:1.2574   1st Qu.:-0.639500   Up  :648
##  Median :1.4229   Median : 0.038500
##  Mean   :1.4783   Mean   : 0.003138
##  3rd Qu.:1.6417   3rd Qu.: 0.596750
##  Max.   :3.1525   Max.   : 5.733000
```

```
train = (Year < 2005);
```

Which of the following is true regarding `train`?

- A). It contains the row indexes for which 'Year' is smaller than 2005.
- B). It contains 1250 True/False values.
- C). It contains True/False values, but the number of values is clearly less than 1250.

answer: B)

7. In ridge regression, the objective function is

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=0}^{p}\beta_j^2\,.$$

The above statement is

- A) Correct.
- B) Not correct, because we should have L1 panalty instead of the L2 penalty here.
- C) Not correct, because we do not panalize the β_0 term.

Answer: C).

8. Which of the following are/is NOT correct about adjusted R-squares.

i) Adjusted R-squares is usually smaller than R-squares

ii) Adjusted R-squares takes values only between 0 and 1.

iii) Adjusted R-squares can be used to compare models of different sizes

- A) i), ii), iii)
- B) i), iii)
- C) i), ii)
- D) i)
- E) ii)
- F) iii)
- G) There is no incorrect statement among i), ii, iii).

Answer: E) Adjusted R-squares can take negative values. Just take a look at the formula.

9. In the AIC criterion, we have

$$AIC=\frac{1}{n\hat{\sigma}^2}\left(RSS+2d\hat{\sigma}^2\right)\,.$$

When d is smaller than than the full model size p,

- A) The $\hat{\sigma}^2$ in AIC is got by using the all p predictors.
- B) The $\hat{\sigma}^2$ in AIC is got by using only d predictors.
- C) Either choice is good.

answer: A).

10.

What characteristic of the data is most likely to have led to the warming message of the following logistic regression analysis?

```
x = c(1, 1.5, 2, 5, 5.5, 6); y = c(0, 0, 0, 1, 1, 1)
lr.fit = glm(y~x, family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

- A) The two classes are not separated well.
- B) The two classes are well separated.
- C) There is nothing special about the data. Logistic regression is a very restrictive method, and does not apply well to many scenarios anyway.

solution: B)

11.

A recent survey conducted by the personnel manager of a major enterprise resources planning (ERP) company showed that 35% of the employees were dissatisfied with their salary, 80% were satisfied with their work assignments, 15% were dissatisfied with their work hours, 17% were dissatisfied with both their salary and work assignments, and 8% were dissatisfied with both their work assignments and work hours. What is the percentage of employees who are satisfied with both their salary and work assignments?

- A) 0.38
- B) 0.02
- C) 0.62
- D) 0.52
- E) None of the above

Answer: : C) $1 - (0.35 + 0.2 - 0.17) = 0.62$ It is tempting to do $1 - (0.35 + 0.8 - 0.17) = 0.02$.

12.

Let Y be the response variable and X be the explanatory variable. Suppose that correlation between X and Y is 0.5, the mean of X is 3.0, the mean of Y is 1.5, the variance of X is 25.0, and the variance of Y is 16.0. Let b_0 be the intercept of the least squares regression line, and b_1 be the slope. What is $b_0 + b_1$?

- A) 0.30
- B) 0.40
- C) 0.60
- D) 0.70
- E) None of the above

Answer: D): $b_1 = 0.5 * \sqrt{16}/\sqrt{25} = 0.4$, and $b_0 = 1.5 - 0.4 * 3.0 = 0.3$

13.

Suppose a dataset has one response variable Y, and four explanatary variables, X_1, X_2, X_3 and X_4. Regress Y on X_1, X_2 and X_3 using the entire dataset, we get the VIF of X_1 equals 11 (i.e. $VIF_1 = 11$). Now we regress Y on X_1, X_2, X_3 and X_4 using the entire dataset. What can we say about VIF_1?

- A) VIF_1 is bigger than 11
- B) VIF_1 is smaller than 11

Answer: A)

14.

Which is the following statements are/is true?

- i) Validation set approach is a special case of cross validation
- ii) Leave-one-out-cross-validation (LOOCV) is a special case of $k-$ fold CV.
- iii) 5-fold CV is always better than 10-fold CV.
- A) i), ii), iii)
- B) i)
- C) ii)
- D) iii)
- E) i), ii)

Answer: C)

15. Since principal component regression (PCR) usually use only a few principal components, it can be thought of as a feature selection method.

- A) True
- B) False

answer: B), because each princiapl component potentially uses all original features.

16. Suppose we want to sample 100 numbers with replacement from $\{1, \cdots, 100\}$. Is the following code correct?

```
sample = sample(1:100, 100)
```

- A) Yes.
- B). No.

solution: B). The correct code shoud be `sample(1:100, 100, rep=T)`.

Problems **17-23** are based on the following dataset.

```
attach(Hitters); summary(Hitters)
```

```
##      AtBat            Hits         HmRun            Runs
##  Min.   : 16.0   Min.   :  1   Min.   : 0.00   Min.   :  0.00
##  1st Qu.:255.2   1st Qu.: 64   1st Qu.: 4.00   1st Qu.: 30.25
##  Median :379.5   Median : 96   Median : 8.00   Median : 48.00
##  Mean   :380.9   Mean   :101   Mean   :10.77   Mean   : 50.91
##  3rd Qu.:512.0   3rd Qu.:137   3rd Qu.:16.00   3rd Qu.: 69.00
##  Max.   :687.0   Max.   :238   Max.   :40.00   Max.   :130.00
##
##       RBI             Walks            Years            CAtBat
```

```
##  Min.   :  0.00   Min.   :  0.00   Min.   : 1.000   Min.   :   19.0
##  1st Qu.: 28.00   1st Qu.: 22.00   1st Qu.: 4.000   1st Qu.:  816.8
##  Median : 44.00   Median : 35.00   Median : 6.000   Median : 1928.0
##  Mean   : 48.03   Mean   : 38.74   Mean   : 7.444   Mean   : 2648.7
##  3rd Qu.: 64.75   3rd Qu.: 53.00   3rd Qu.:11.000   3rd Qu.: 3924.2
##  Max.   :121.00   Max.   :105.00   Max.   :24.000   Max.   :14053.0
##
##      CHits            CHmRun           CRuns             CRBI
##  Min.   :   4.0   Min.   :  0.00   Min.   :   1.0   Min.   :   0.00
##  1st Qu.: 209.0   1st Qu.: 14.00   1st Qu.: 100.2   1st Qu.:  88.75
##  Median : 508.0   Median : 37.50   Median : 247.0   Median : 220.50
##  Mean   : 717.6   Mean   : 69.49   Mean   : 358.8   Mean   : 330.12
##  3rd Qu.:1059.2   3rd Qu.: 90.00   3rd Qu.: 526.2   3rd Qu.: 426.25
##  Max.   :4256.0   Max.   :548.00   Max.   :2165.0   Max.   :1659.00
##
##      CWalks        League  Division    PutOuts          Assists
##  Min.   :   0.00   A:175   E:157    Min.   :   0.0   Min.   :  0.0
##  1st Qu.:  67.25   N:147   W:165    1st Qu.: 109.2   1st Qu.:  7.0
##  Median : 170.50                    Median : 212.0   Median : 39.5
##  Mean   : 260.24                    Mean   : 288.9   Mean   :106.9
##  3rd Qu.: 339.25                    3rd Qu.: 325.0   3rd Qu.:166.0
##  Max.   :1566.00                    Max.   :1378.0   Max.   :492.0
##
##      Errors          Salary       NewLeague
##  Min.   : 0.00   Min.   :  67.5   A:176
##  1st Qu.: 3.00   1st Qu.: 190.0   N:146
##  Median : 6.00   Median : 425.0
##  Mean   : 8.04   Mean   : 535.9
##  3rd Qu.:11.00   3rd Qu.: 750.0
##  Max.   :32.00   Max.   :2460.0
##                  NA's   :59
```

17. How many players in the dataset are NOT missing salary numbers?

- A) 59
- B) 322
- C) None of the above.

answer: C) $322 - 59 = 263$.

18. To predict the players' salary based on this data set, we first delete the rows with missing values:

```
Hitters = na.omit(Hitters)
```

Choose the first 150 rows as training set, and let `Hitters.test` be the test set that include the rest of the rows in `Hitters`. How to get `Hitters.train`?

- A) Hitters.train = Hitters[1:150,]
- B) Hitters.train = Hitters[,1:150]
- C) Hitters.train = Hitters[sample(1:nrow(Hitters), 150),]
- D) Hitters.train = Hitters[,sample(1:nrow(Hitters), 150)]

answer: A)

19.

To implement best subset selection,

```
library(leaps)
regfit.best = regsubsets(Salary~., data = Hitters.train, nvmax = 19)
summary(regfit.best)
```

```
## Subset selection object
## Call: regsubsets.formula(Salary ~ ., data = Hitters.train, nvmax = 19)
## 19 Variables  (and intercept)
##            Forced in Forced out
## AtBat          FALSE      FALSE
## Hits           FALSE      FALSE
## HmRun          FALSE      FALSE
## Runs           FALSE      FALSE
## RBI            FALSE      FALSE
## Walks          FALSE      FALSE
## Years          FALSE      FALSE
## CAtBat         FALSE      FALSE
## CHits          FALSE      FALSE
## CHmRun         FALSE      FALSE
## CRuns          FALSE      FALSE
## CRBI           FALSE      FALSE
## CWalks         FALSE      FALSE
## LeagueN        FALSE      FALSE
## DivisionW      FALSE      FALSE
## PutOuts        FALSE      FALSE
## Assists        FALSE      FALSE
## Errors         FALSE      FALSE
## NewLeagueN     FALSE      FALSE
## 1 subsets of each size up to 19
## Selection Algorithm: exhaustive
##           AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns
## 1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "
## 2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   "*"    " "
## 3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "
## 4  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "    " "   "*"    " "
## 5  ( 1 )  " "   "*"  " "   " "  " " " "   " "   "*"    "*"   "*"    " "
## 6  ( 1 )  " "   "*"  " "   " "  " " " "   " "   "*"    "*"   "*"    " "
## 7  ( 1 )  " "   "*"  " "   " "  "*" " "   " "   "*"    "*"   "*"    " "
## 8  ( 1 )  " "   "*"  " "   " "  "*" "*"   " "   "*"    "*"   "*"    " "
## 9  ( 1 )  "*"   "*"  "*"   " "  " " "*"   " "   "*"    "*"   "*"    " "
## 10  ( 1 ) "*"   "*"  " "   " "  "*" "*"   " "   "*"    " "   " "    "*"
## 11  ( 1 ) "*"   "*"  "*"   " "  " " "*"   " "   "*"    "*"   "*"    " "
## 12  ( 1 ) "*"   "*"  "*"   " "  " " "*"   " "   "*"    "*"   "*"    " "
## 13  ( 1 ) "*"   "*"  " "   " "  "*" "*"   "*"   "*"    "*"   "*"    " "
## 14  ( 1 ) "*"   "*"  " "   " "  "*" "*"   "*"   "*"    "*"   "*"    " "
## 15  ( 1 ) "*"   "*"  " "   " "  "*" "*"   "*"   "*"    "*"   "*"    "*"
## 16  ( 1 ) "*"   "*"  " "   " "  "*" "*"   "*"   "*"    "*"   "*"    "*"
## 17  ( 1 ) "*"   "*"  "*"   " "  "*" "*"   "*"   "*"    "*"   "*"    "*"
```

```
## 18  ( 1 ) "*"   "*"  "*"   " "  "*" "*"   "*"   "*"    "*"   "*"    "*"
## 19  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   "*"   "*"    "*"   "*"    "*"
##           CRBI CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
## 1  ( 1 )  "*"  " "    " "     " "       " "     " "     " "    " "
## 2  ( 1 )  " "  " "    " "     " "       " "     " "     " "    " "
## 3  ( 1 )  "*"  " "    " "     " "       "*"     " "     " "    " "
## 4  ( 1 )  " "  " "    " "     " "       "*"     " "     " "    " "
## 5  ( 1 )  " "  " "    " "     " "       "*"     " "     " "    " "
## 6  ( 1 )  " "  " "    " "     "*"       "*"     " "     " "    " "
## 7  ( 1 )  " "  " "    " "     "*"       "*"     " "     " "    " "
## 8  ( 1 )  " "  " "    " "     "*"       "*"     " "     " "    " "
## 9  ( 1 )  " "  " "    " "     "*"       "*"     " "     " "    " "
## 10  ( 1 ) "*"  "*"    " "     "*"       "*"     " "     " "    " "
## 11  ( 1 ) " "  " "    " "     "*"       "*"     "*"     "*"    " "
## 12  ( 1 ) " "  "*"    " "     "*"       "*"     "*"     "*"    " "
## 13  ( 1 ) " "  "*"    " "     "*"       "*"     "*"     "*"    " "
## 14  ( 1 ) "*"  "*"    " "     "*"       "*"     "*"     "*"    " "
## 15  ( 1 ) "*"  "*"    " "     "*"       "*"     "*"     "*"    " "
## 16  ( 1 ) "*"  "*"    " "     "*"       "*"     "*"     "*"    "*"
## 17  ( 1 ) "*"  "*"    " "     "*"       "*"     "*"     "*"    "*"
## 18  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
## 19  ( 1 ) "*"  "*"    "*"     "*"       "*"     "*"     "*"    "*"
```

The above summary indicates, the best model of size 1 includes the variable

- A) AtBat
- B) CRBI
- C) Years
- D) CHmRun

answer: B), the one has a star in the row indexed 1.

20.

```
summary(regfit.best)$adjr2
```

```
##  [1] 0.4217242 0.5649680 0.6052412 0.6274570 0.6547240 0.6652501 0.6712912
##  [8] 0.6780375 0.6828723 0.6847795 0.6867436 0.6878499 0.6893731 0.6880220
## [15] 0.6864356 0.6850644 0.6830696 0.6808279 0.6783755
```

By adjusted R^2, which model size do you pick?

- A)1
- B) 2
- C) 11
- D) 12
- E) 13
- F) 14

answer: E) we choose the model with the largest adjusted R^2.

21. Now we apply penalized regression

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loading required package: foreach
```

```
## Loaded glmnet 2.0-16
```

```
x.train = model.matrix(Salary ~., Hitters.train)[, -1]
y.train = Hitters.train$Salary
x.test = model.matrix(Salary ~., Hitters.test)[, -1]
y.test = Hitters.test$Salary
grid = 10^seq(10, -3, length = 200)
mod1 = glmnet(x.train, y.train, alpha = 0, lambda = grid)
```

What is the smallest value of λ in the grid?

- A) 10^{10}
- B) 10
- C) -3
- D) 10^{-3}
- E) 0

answer: D).

22. The $\alpha = 0$ in the above `glmnet` function means we are applying

- A) LASSO regression
- B) ridge regression
- C) It could be either LASSO or ridge regression; we do not have enough information to figure this out.

answer: B)

23. From the following output, which λ is a better choice?

```
pred10 = predict(mod1, s = 10, newx=x.test)
mean((pred10 - y.test)^2)
```

```
## [1] 143991.4
```

```
pred100 = predict(mod1, s = 100, newx=x.test)
mean((pred100 - y.test)^2)
```

```
## [1] 136146.5
```

- A) $\lambda = 10$
- B) $\lambda = 100$

answer: B), because $\lambda = 100$ has a smaller test MSE.

24. If a decision tree (defined as in `ISLR`) partitions the feature space into regions $R_1 \cdots, R_J$, can any of these regions be a ball?

- A) Yes.
- B) No.

answer: B).

25. In the bagging, random forests and boosting algorithms, which involve boostrap samples?

- A) all three
- B) bagging and random forest
- C) bagging and boosting
- D) random forests and boosting

answer: B)

Questions **26-30** are short answer questions.

26.

```
set.seed(2); x = seq(1:4); y = x + rnorm(4, 0, 0.5);
linear_model = lm(y~x); summary(linear_model); hatvalues(linear_model)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##       1       2       3       4
## -0.3639  0.1418  0.8082 -0.5861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -0.1197     0.9284  -0.129   0.9092
## x             1.0351     0.3390   3.054   0.0926 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.758 on 2 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.7351
## F-statistic: 9.324 on 1 and 2 DF,  p-value: 0.0926
```

```
##   1   2   3   4
## 0.7 0.3 0.3 0.7
```

i) Compute the LOOCV from the R output above. Keep two decimal places for the answer. ii) If we were to eliminate "hatvalues(linear_model)" from the above R code, can you still compute LOOCV in this example, why?

Solution: i) LOOCV for the linear regression has a special formula:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1-h_i}\right)^2$$

$CV_{(4)} = \frac{1}{4}\left((-0.3639/(1-0.7))^2 + (0.1418/(1-0.3))^2 + (0.8082/(1-0.3))^2 + (-0.5861/(1-0.7))^2\right) = 1.67$

ii) We can compute the h_i's based on the x's. So we can still compute the $CV_{(4)}$.

For ii) stating the idea that h_i's could be obtained from x's is ok.

1 point for i) and 1 point for ii)

27. When Robert applies LASSO in linear regression (number of features $p = 3$) with $\lambda = 100$ (where λ is the penalty parameter), the estimated coefficients are $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (10, 5, 0)$. When he applies LASSO again on the same dataset with $\lambda = 105$, he finds the estimated coefficients to be $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (13, 5, 0)$. Without looking at Robert's data or his analysis, his superviser says that at least one of these LASSO regressions cannot be correct. Can you tell the rationale behind the superviser's judgement.

solution: $100 < 105$, but $10 + 5 + 0 < 13 + 5 + 0$.

As long as one can correctly argue as λ increase, the L_1 norm should decrease, then one will have full credit. If one argue that every coordinate of $\hat{\beta}$ should decrease, or similarly, in a wrong way of concluding L_1 should decrease, then one will receive only 1 point.

28. Suppose your manager knows a little (not much) about decision trees. When you show him your work, he finds out the `wage` variable appears in three different nodes. He is unhappy about it, and tell you every variable should appear at most one node in a decision tree. Do you agree with his comment? If not, how will you try to convince your manager that requiring each variable to appear at at most once is not a good idea?

solution: We do not agree with the manager. In decision trees, if we restrict each variable to appear at most once, we will pose severe limitation on the ways the feature space can be partitioned (a high bias situation).

Alternatively, tree constructing is to minimize RSS, restricting each variable to appear at most once will make the tree less able to do that.

Full credits for reasonable argument.

29. In random forest algorithms, we restrict our attention to randomly selected m out of p features in each split. Now we change this procedure to restriction to the first m features (i.e., $X_1, \cdots, X_m$) in every split. Do you expect the new procedure to work well? And why?

answer: The procedure is unlikely to work well. Random selection of m features in each split `decorrelates` the trees, but the new procedure does not.

Alternatively, the new procedure may miss important variables.

Full credits for expressing the idea of the new procedure no longer serving the purpose of decorrelation or missing important variables.

30. In the *forward stepwise selection* algorithm, C_p and BIC criterions among others can be used to select model sizes. When the sample size is 100, argue why the BIC criterion selects smaller models compared to C_p.

answer:

Full credits for showing that BIC poses a heavier penalty on model size compared to C_p when n is greater than 100.