

DSC 324/424 Assignment 1

Answer all questions as completely as you can. Submit your answers to the following prompts in a single document to the Assignment 1 submission folder by the due date. **Make sure your name is clearly written at the top of your document.** Along with this document, submit a .R file with the code you used to complete the assignment. You may do any required math on paper, as long as you include *clear* screenshots or scans of your work in your final document.

The file matrix_snippets.R on d2l has some helpful code for problems 1, 2, and 3.

1. Vector multiplication (dot product): Using a dot product, calculate the variance of the following variable values by hand. Show your work and include your final answer. Use the matrix operators in R to calculate the same and confirm your result. Include a screenshot of your code and output.

[8 2 2 3]

Solution:

Handwritten solution for calculating the variance of the vector $X = [8, 2, 2, 3]$.

1. $X = [8, 2, 2, 3]$

$\bar{X} = \frac{8+2+2+3}{4} \Rightarrow \bar{X} = 3.75$

Variance $\Rightarrow \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

$= \frac{(8-3.75)^2 + (2-3.75)^2 + (2-3.75)^2 + (3-3.75)^2}{3}$

$= \frac{18.0625 + 3.0625 + 3.0625 + 0.5625}{3}$

$= 8.25$

```

2 #1
3 x = c(8,2,2,3)
4 x
5 var(x)
6
7 x = mean(x)
8 x
9
10 Variance <- ((x-3.75)%*(x-3.75))/3
11 Variance
12
13 #2
14
15 A<- matrix(c(4,3, 2,7),nrow= 2,ncol= 2, byrow = T)
16 A
17 cov(A)
18 a = t(A)
19 covar<- a%*%A
20 covar
21

```

```

21:1 (Top Level) : R Script :
Console Terminal Jobs x
> x = c(8,2,2,3)
> x
[1] 8 2 2 3
> var(x)
[1] 8.25
> x = mean(x)
> x
[1] 3.75
> Variance <- ((x-3.75)%*(x-3.75))/3
> Variance
[1,]
[1,] 8.25
> A<- matrix(c(4,3, 2,7),nrow= 2,ncol= 2, byrow = T)
> A
[1,] [1,] [2,]
[1,] 4 3
[2,] 2 7
> cov(A)

```

2. Matrix Multiplication: Calculate the covariance matrix of the following stock portfolios by hand. Show your work and include your final matrix:

Covariance of all variables in a dataset = $X^T X$

	Intel	AMD
Jim	4	3
Kate	2	7

Solutions:

2. Covariance.

$$\begin{matrix} 4 & 3 \\ 2 & 7 \end{matrix}$$

$$= (4, 3) (2, 7)$$

$$\bar{x} = \frac{4+2}{2} = 3.$$

$$\bar{y} = \frac{3+7}{2} = 5.$$

$$\begin{aligned} \text{Cov}(x, y) &= \frac{\sum_{i=1}^k (x_i - \bar{x})(y_i - \bar{y})}{k-1} \\ &= \frac{(4-3)(3-5) + (2-3)(7-5)}{2-1} \\ &= \frac{1(-2) + (-1)(2)}{1} \\ &= -4. \end{aligned}$$

$$\begin{aligned} \text{Variance of } x &= \frac{\sum_{i=1}^k (x_i - \bar{x})^2}{k-1} \\ &= \frac{(4-3)^2 + (2-3)^2}{2-1} \\ &= 2. \end{aligned}$$

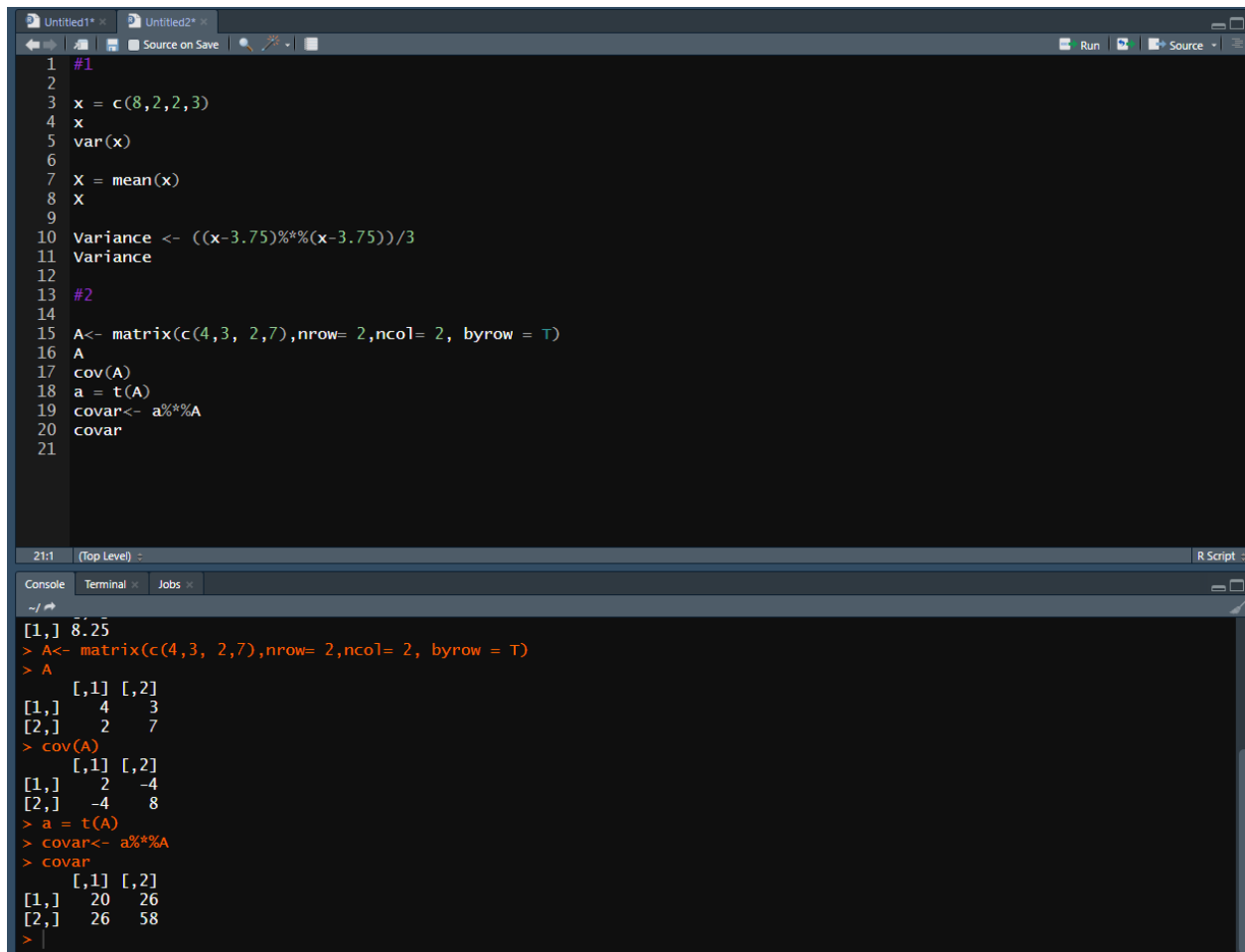
$$\begin{aligned} \text{Variance of } y &= \frac{\sum_{i=1}^k (y_i - \bar{y})^2}{k-1} \\ &= \frac{(3-5)^2 + (7-5)^2}{2-1} \\ &= 8. \end{aligned}$$

Cov Matrix =

$$\begin{bmatrix} \text{Var}(x) & \text{Cov}(x, y) \\ \text{Cov}(x, y) & \text{Var}(y) \end{bmatrix}$$

$$= \begin{bmatrix} 2 & -4 \\ -4 & 8 \end{bmatrix}$$

Use the matrix operators in R to calculate the same and confirm your result. Include a screenshot of your code and output. Is there anything notable about this matrix?



```
1 #1
2
3 x = c(8,2,2,3)
4 x
5 var(x)
6
7 X = mean(x)
8 X
9
10 Variance <- ((x-3.75)%*(x-3.75))/3
11 Variance
12
13 #2
14
15 A<- matrix(c(4,3, 2,7),nrow= 2,ncol= 2, byrow = T)
16 A
17 cov(A)
18 a = t(A)
19 covar<- a%*%A
20 covar
21
```

```
[1,] 8.25
> A<- matrix(c(4,3, 2,7),nrow= 2,ncol= 2, byrow = T)
> A
      [,1] [,2]
[1,]    4    3
[2,]    2    7
> cov(A)
      [,1] [,2]
[1,]    2   -4
[2,]   -4    8
> a = t(A)
> covar<- a%*%A
> covar
      [,1] [,2]
[1,]   20   26
[2,]   26   58
>
```

4. Variable selection and regression: Use the *housing.csv* dataset to answer the following prompts concerning automatic variable selection techniques.

This Housing dataset contains housing values in the suburbs of Boston. A brief description of the variables can be found below. A more detailed explanation of the dataset can be found at the UCI machine learning repository <http://archive.ics.uci.edu/ml/datasets/Housing>

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940

- DIS: weighted distances to five Boston employment centers
 - RAD: index of accessibility to radial highways
 - TAX: full-value property-tax rate per \$10,000
 - PTRATIO: pupil-teacher ratio by town
 - LSTAT: % lower status of the population
 - MEDV: Median value of owner-occupied homes in \$1000's (output variable)
- a. Fit a linear regression model of CRIM based on all the other variables and include a screenshot of the statistics generated by the `lm()` function.

```

> model1 = lm(CRIM ~ ., data=housing)
> summary(model1)

Call:
lm(formula = CRIM ~ ., data = housing)

Residuals:
    Min       1Q   Median       3Q      Max
-8.736 -2.269 -0.404  1.098  72.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.399431   8.915586   2.176  0.030200 *
      ZN         0.051780   0.023656   2.189  0.029233 *
    INDUS      -0.042325   0.104388  -0.405  0.685373
    CHAS      -0.969182   1.445497  -0.670  0.502971
    NOX      -12.393952   6.758412  -1.834  0.067485 .
      RM         0.719517   0.805262   0.894  0.372165
      AGE         0.003184   0.022673   0.140  0.888383
      DIS      -1.218972   0.362310  -3.364  0.000848 ***
      RAD         0.679897   0.106355   6.393  4.95e-10 ***
      TAX        -0.006016   0.006253  -0.962  0.336684
    PTRATIO    -0.371992   0.237939  -1.563  0.118822
    LSTAT      0.028566   0.098464   0.290  0.771894
    MEDV      -0.292519   0.076829  -3.807  0.000165 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.995 on 367 degrees of freedom
Multiple R-squared:  0.4257,    Adjusted R-squared:  0.4069
F-statistic: 22.67 on 12 and 367 DF, p-value: < 2.2e-16

```

- b. Evaluate this model's performance using the evaluation metrics you've learned for multiple linear regression. In your evaluation, make a case for whether or not this model is performing sufficiently well in the context of this dataset.

Here, Model has adjusted R-square 0.4069. After observing this we can say that model passes the hypothesis test as P value is less than 0.05 but there are variables does not pass the t-test.

There are certain variables that seem statistically significant such as DIS ,RAD, MEDV,ZN. Also the NOX variable can be considered since it has its P value 0.06.

Hence we can accept the Alternate hypothesis and reject the null hypothesis

- c. Fit two more linear regressions with the same variables, one using the forward selection and one using the backwards elimination variable selection techniques. Include screenshot of both models' statistics as you did in part a.

FORWARD SELECTION

```
Console Terminal Jobs
~/
> # First we do a forward search
> housingForward = step(none, scope = list(lower=none, upper=all),
+                       direction="forward", trace=F)
> summary(housingForward)

Call:
lm(formula = CRIM ~ RAD + MEDV + PTRATIO + DIS + NOX + ZN, data = housing)

Residuals:
    Min       1Q   Median       3Q      Max
-8.856 -2.189 -0.335  1.006 73.761

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.74595    7.44076   3.057  0.00240 **
RAD           0.60002    0.05927  10.123 < 2e-16 ***
MEDV        -0.25471    0.05142  -4.953 1.11e-06 ***
PTRATIO     -0.41246    0.23240  -1.775  0.07675 .
DIS         -1.13649    0.32103  -3.540  0.00045 ***
NOX        -13.92060    6.14197  -2.266  0.02399 *
ZN           0.04852    0.02253   2.153  0.03195 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.971 on 373 degrees of freedom
Multiple R-squared:  0.4203,    Adjusted R-squared:  0.4109
F-statistic: 45.07 on 6 and 373 DF,  p-value: < 2.2e-16
```

BACKWARD SELECTION

```
Console Terminal Jobs
~/
> housingBackward = step(all, direction="backward", trace=F)
> summary(housingBackward)

Call:
lm(formula = CRIM ~ ZN + NOX + DIS + RAD + PTRATIO + MEDV, data = housing)

Residuals:
    Min       1Q   Median       3Q      Max
-8.856 -2.189 -0.335  1.006 73.761

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.74595    7.44076   3.057  0.00240 **
ZN           0.04852    0.02253   2.153  0.03195 *
NOX        -13.92060    6.14197  -2.266  0.02399 *
DIS         -1.13649    0.32103  -3.540  0.00045 ***
RAD           0.60002    0.05927  10.123 < 2e-16 ***
PTRATIO     -0.41246    0.23240  -1.775  0.07675 .
MEDV        -0.25471    0.05142  -4.953 1.11e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.971 on 373 degrees of freedom
Multiple R-squared:  0.4203,    Adjusted R-squared:  0.4109
F-statistic: 45.07 on 6 and 373 DF,  p-value: < 2.2e-16

> |
```

- d. Compare the variables the two selection methods decided to keep. Is there any consensus? Were there any variables kept that were not deemed statistically significant by the model in part a?

The adjusted R-square for both forward and backward model is 0.4109

After performing both forward and backward selection, We see that both the models seem to be exactly same with same set of variables considered.

PTRATIO and NOX are the variables present in both selection methods which are not in the complete linear model built earlier and their P-value has changed after removing some variable from the model.

NOX now has 0.02 P-value which is 0.06 in complete model as same PTRATIO has P-value of 0.07 which was 0.1 in complete model.

RAD, ZN, DIS, MEDV are the independent variable which are statistically significant in all models.

e. Compare the performance of all three models and make a case for which is most appropriate to use for predicting crime rates. Consider whether or not the chosen variables are appropriate predictors for this purpose. Also consider whether or not the benefits of having a simpler model justify accepting slightly worse performance than a more complex model.

After observing and reviewing all the three models we can say that both the forward and backward selection methods gave us significant models with less number of variables. Nonetheless all the three models are having their Adjusted R Squared values almost equal.

Since both the forward and backward selection models gave us the same results with less number of significant variables thus can be a good parsimonious model with adjusted R square values 0.4109 which is not that bad. We can say that both the forward and backward selection models can be used for Predictions as they make models less complex with some good performance than the complete model which performs almost similar but with more number of less significant variables.