MAST9014: Introduction to Statistical Learning

Assignment 1 Solutions

1. Suppose that A is a symmetric matrix with $A^k = A^{k+1}$ for some integer $k \geq 1$. Show that A is idempotent.

   Solution: Let λ be an eigenvalue of A with eigenvector $\mathbf{x}$. Then

   $$\begin{aligned}\lambda^k \mathbf{x} &= A^k \mathbf{x} \\ &= A^{k+1}\mathbf{x} \\ &= \lambda^{k+1}\mathbf{x}.\end{aligned}$$

   Hence $\lambda^k(1-\lambda) = 0$, so $\lambda = 0$ or 1. Therefore all eigenvalues of A are 0 or 1.

   [1]

   (NB: can get to this via diagonalisation and $PD^kP^T = A^k = A^{k+1} = PD^{k+1}P^T$ so $D^k = D^{k+1}$ giving eigenvalues all 0 or 1.)

   Now diagonalise A and write it as

   $$A = PDP^T.$$

   Then

   $$\begin{aligned}A^2 &= PDP^TPDP^T \\ &= PD^2P^T \\ &= PDP^T = A\end{aligned}$$

   since D has only 0 or 1 on the diagonal and hence $D^2 = D$.

   [1]

2. Let $A_1, A_2, \ldots, A_m$ be a set of symmetric $k \times k$ matrices. Suppose that there exists an orthogonal matrix P such that P^TA_iP is diagonal for all i. Show that $A_iA_j = A_jA_i$ for every pair $i, j = 1, 2, \ldots, m$.

   Solution:

   $$\begin{aligned}P^TA_iA_jP &= \left(P^TA_iP\right)\left(P^TA_jP\right) \\ &= \left(P^TA_jP\right)\left(P^TA_iP\right) \\ &= P^TA_jA_iP.\end{aligned}$$

   Premultiply by P and postmultiply by P^T to get $A_iA_j = A_jA_i$.

   [2]

3. Show that for any random vector $\mathbf{y}$ and compatible matrix A, we have var $A\mathbf{y} = A(\text{var } \mathbf{y})A^T$.

   Solution: Let $\boldsymbol{\mu} = E[\mathbf{y}]$. From the definition,

   $$\begin{aligned}\text{var } A\mathbf{y} &= E[(A\mathbf{y} - A\boldsymbol{\mu})(A\mathbf{y} - A\boldsymbol{\mu})^T] \\ &= E[A(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^TA^T] \\ &= A\,E[(\mathbf{y}-\boldsymbol{\mu})(\mathbf{y}-\boldsymbol{\mu})^T]\,A^T \\ &= A(\text{var } \mathbf{y})A^T.\end{aligned}$$

   [2]

4. Let $\mathbf{y}$ be a 3-dimensional multivariate normal random vector with mean and variance

$$\boldsymbol{\mu} = \begin{bmatrix} 3 \\ 0 \\ -2 \end{bmatrix}, \quad V = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let

$$A = \frac{1}{10} \begin{bmatrix} 4 & -2 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 10 \end{bmatrix}.$$

(a) Describe the distribution of $A\mathbf{y}$.
Solution: $A\mathbf{y} \sim MVN(A\boldsymbol{\mu}, AVA^T)$.

```
mu <- c(3,0,-2)
V <- diag(c(2,2,1))
A <- matrix(c(4,-2,0,-2,1,0,0,0,10)/10,3,3)
A%*%mu
```

```
##      [,1]
## [1,]  1.2
## [2,] -0.6
## [3,] -2.0
```

```
A%*%V%*%t(A)
```

```
##      [,1] [,2] [,3]
## [1,]  0.4 -0.2    0
## [2,] -0.2  0.1    0
## [3,]  0.0  0.0    1
```

[2]

(b) Find $E[\mathbf{y}^T A\mathbf{y}]$.

```
sum(diag(A%*%V)) + t(mu)%*%A%*%mu
```

```
##      [,1]
## [1,]  9.6
```

[2]

(c) Describe the distribution of $\mathbf{y}^T A\mathbf{y}$.

```
A%*%V
```

```
##      [,1] [,2] [,3]
## [1,]  0.8 -0.4    0
## [2,] -0.4  0.2    0
## [3,]  0.0  0.0    1
```

```
A%*%V%*%A%*%V
```

```
##      [,1] [,2] [,3]
## [1,]  0.8 -0.4    0
## [2,] -0.4  0.2    0
## [3,]  0.0  0.0    1
```

```
library(Matrix)
rankMatrix(A%*%V)[1]
```

```
## [1] 2
```

AV is idempotent and has rank 2.
Hence $\mathbf{y}^T A\mathbf{y}$ has a non-central χ^2 distribution with 2 degrees of freedom and noncentrality parameter:

```
t(mu)%*%A%*%mu/2
```

```
##      [,1]
## [1,]  3.8
```

[2]

(d) Find all linear combinations of $\mathbf{y}$ elements which are independent of $\mathbf{y}^T A\mathbf{y}$.
Solution: These linear combinations are of the form $\mathbf{c}^T\mathbf{y}$, where $\mathbf{c}^T VA = 0$. This gives

$$\begin{aligned} 0.8c_1 - 0.4c_2 &= 0 \qquad &\Rightarrow \quad c_2 = 2c_1 \\ c_3 &= 0 \end{aligned}$$

Thus all linear combinations of $\mathbf{y}$ elements that are independent of $\mathbf{y}^T A\mathbf{y}$ are multiples of $y_1 + 2y_2$.

[2]

5. The table below shows prices in US cents per pound received by fishermen and vessel owners for various species of fish and shellfish in 1970 and 1980. (Taken from Moore & McCabe, Introduction to the Practice of Statistics, 1989.)

Type of fish	Price (1970)	Price (1980)
Cod	13.1	27.3
Flounder	15.3	42.4
Haddock	25.8	38.7
Menhaden	1.8	4.5
Ocean perch	4.9	23.0
Salmon, chinook	55.4	166.3
Salmon, coho	39.3	109.7
Tuna, albacore	26.7	80.1
Clams, soft-shelled	47.5	150.7
Clams, blue hard-shelled	6.6	20.3
Lobsters, american	94.7	189.7
Oysters, eastern	61.1	131.3
Sea scallops	135.6	404.2
Shrimp	47.6	149.0

We will model this data using a linear model.

(a) The linear model is of the form $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Write down the matrices and vectors involved in this equation.
Solution:

$$\mathbf{y} = \begin{bmatrix} 27.3 \\ 42.4 \\ 38.7 \\ 4.5 \\ 23.0 \\ 166.3 \\ 109.7 \\ 80.1 \\ 150.7 \\ 20.3 \\ 189.7 \\ 131.3 \\ 404.2 \\ 149.0 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 13.1 \\ 1 & 15.3 \\ 1 & 25.8 \\ 1 & 1.8 \\ 1 & 4.9 \\ 1 & 55.4 \\ 1 & 39.3 \\ 1 & 26.7 \\ 1 & 47.5 \\ 1 & 6.6 \\ 1 & 94.7 \\ 1 & 61.1 \\ 1 & 135.6 \\ 1 & 47.6 \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \end{bmatrix}.$$

[2]

(b) Find the least squares estimators of the parameters.

Solution:

```
n <- 14
p <- 2
y <- c(27.3,42.4,38.7,4.5,23.0,166.3,109.7,80.1,150.7,20.3,189.7,131.3,404.2,149.0)
X <- matrix(c(rep(1,n),13.1,15.3,25.8,1.8,4.9,55.4,39.3,26.7,47.5,6.6,94.7,61.1,
                       135.6,47.6),n,p)
( b <- solve(t(X)%*%X, t(X)%*%y) )
```

```
##            [,1]
## [1,] -1.233836
## [2,]  2.701553
```

[2]

(c) Calculate the sample variance s^2.

Solution:

```
e <- y-X%*%b
( s2 <- sum(e^2)/(n-p) )
```

```
## [1] 777.1528
```

[2]

(d) A fisherman sold ocean trout for 18c/pound in 1970. Predict the price for ocean trout in 1980.

Solution:

```
c(1,18)%*%b
```

```
##           [,1]
## [1,] 47.39412
```

[2]

(e) Calculate the standardised residual for sea scallops.

Solution:

```
H <- X %*% solve(t(X)%*%X) %*% t(X)
z <- e / sqrt(s2 * (1 - diag(H)))
z[13]

## [1] 2.104999
```

[2]

(f) Calculate the Cook's distance for sea scallops.

Solution:

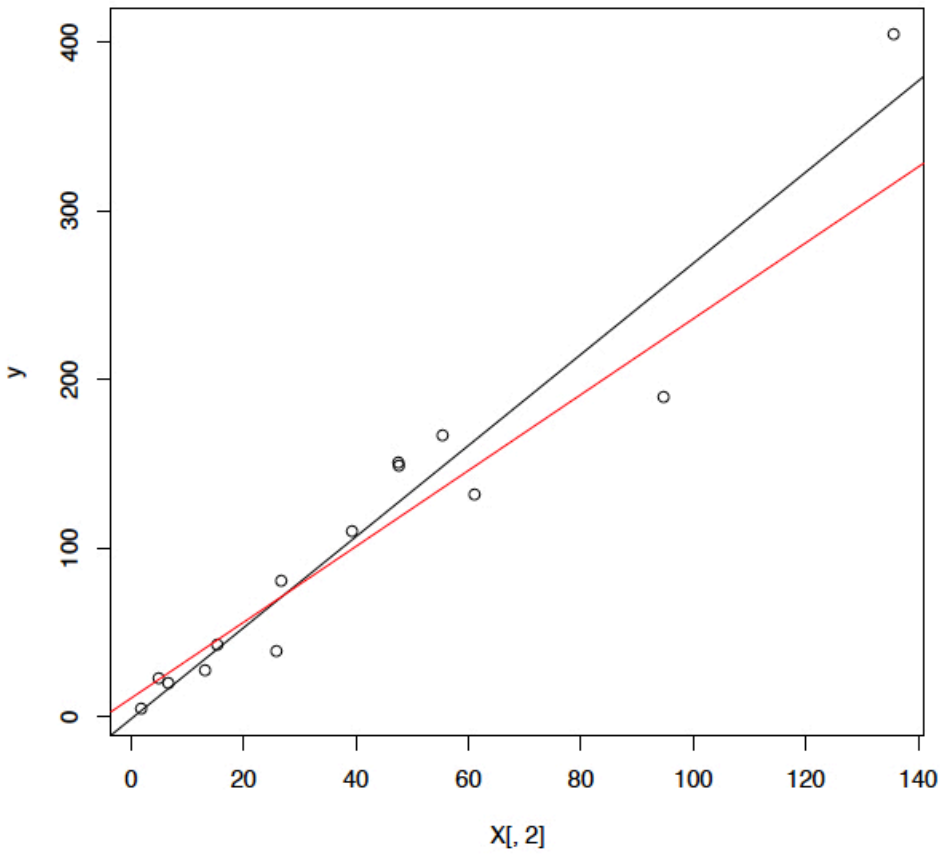
```
D <- 1/p * z^2 * diag(H) / (1-diag(H))
D[13]

## [1] 2.774008
```

[2]

(g) Does sea scallops fit the linear model? Justify your argument.

Solution:

```
b13 <- lm(y[-13] ~ X[-13,2])$coefficients
plot(X[,2],y)
abline(b[1],b[2])
abline(b13[1],b13[2], col="red")
```



The Cook's distance certainly indicates it should be of some concern; however looking at the plot, it seems that the fit is actually okay. There is considerable evidence for heteroskedasticity — the variance increases with x (the design variable). Sea scallops has (by far) the largest x and so may be prone to a larger variance than the remaining points.

[1]

The high Cook's distance therefore comes primarily from a very high leverage, rather than a bad fit to the model.

[1]