

STAT 331 - Applied Linear Models

Cameron Roopnarine

Last updated: September 21, 2020

Contents

Contents	1
0.1 Multiple Linear Regression (MLR)	10

LECTURE 1 | 2020-09-08

Regression model infers the relationship between:

- Response (dependent) variable: variable of primary interest, denoted by a capital letter such as Y .
- Explanatory (independent) variables: (covariates, predictors, features) variables that potentially impact response, denoted (x_1, x_2, \dots, x_p) .

Alligator data:

- Y : length (m)
- x_1 : male/female (categorical, 0 or 1)

Mass in stomach:

- x_2 : fish
- x_3 : invertebrates
- x_4 : reptiles
- x_5 : birds
- x_6, \dots, x_p : other variables

We imagine we can explain Y in terms of (x_1, \dots, x_p) using some function so that $Y = f(x_1, \dots, x_p)$.

In this course, we will be looking at linear models.

The Linear regression model assumes that

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

- Y = value of response
- x_1, \dots, x_p = values of p explanatory variables (assumed to be fixed constants)
- $\beta_0, \beta_1, \dots, \beta_p$ = model parameters
 - β_0 = intercept, expected value of Y when all $x_j = 0$.
 - β_1, \dots, β_p all quantify effect on x_j on Y , $j = 1, \dots, p$
 - ε = random error

A good quote:

“All models are wrong, but some are useful.”

Assume $\varepsilon \sim N(0, \sigma^2)$. In general, the model will not perfectly explain the data.

Q: What is the distribution of Y under these assumptions?

We know:

- $E[Y] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$, and
- $\text{Var}[Y] = \text{Var}[\varepsilon] = \sigma^2$.

Therefore,

$$Y \sim N(\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p, \sigma^2)$$

LECTURE 2 | 2020-09-09

A linear model with response variable (Y) and *one* explanatory variable (x) is called a **simple linear regression**; that is,

$$\bar{Y} = \beta_0 + \beta_1 x + \varepsilon$$

Data consists of pairs (x_i, y_i) where $i = 1, \dots, n$.

Before fitting any model, we might

- make a scatterplot to visualize if there is a linear relationship between x and y
- calculate *correlation*

If X and Y are random variables, then

$$\rho = \text{Corr}[X, Y] = \frac{\text{Cov}[X, Y]}{\text{Sd}[X] \text{Sd}[Y]}$$

Based on (x_i, y_i) we can estimate the sample correlation:

$$\begin{aligned} r &= \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} \end{aligned}$$

The sample correlation measures the strength and direction of the *linear* relationship between X and Y .

- $|r| \approx 1$ strong linear relationship
- $|r| \approx 0$ lack of linear relationship
- $r > 0$ positive relationship
- $r < 0$ negative relationship
- $-1 \leq r \leq 1$

But does not tell us how to predict Y from X . To do so, we need to estimate β_0 and β_1 .

For data (x_i, y_i) for $i = 1, \dots, n$, the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Assume

$$\varepsilon \stackrel{\text{iid}}{\sim} N(0, \sigma^2)$$

Therefore,

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$$

In other words,

$$E[Y_i] = \mu_i = \beta_0 + \beta_1 x_i \text{ and } \text{Var}[Y_i] = \sigma^2$$

Note that the Y_i 's are independent, but they are *not* independently distributed.

Use the *Least Squares* (LS) to estimate β_0 and β_1 .

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2 = S(\beta_0, \beta_1)$$

LS is equivalent to MLE when ε_i 's are iid and Normal.

Taking partial derivatives:

$$\frac{dS}{d\beta_0} = 2 \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] (-1)$$

$$\frac{dS}{d\beta_1} = 2 \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)] (-x_i)$$

Now,

$$\frac{dS}{d\beta_0} = 0 \iff \sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0 \iff \beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\begin{aligned} \frac{dS}{d\beta_1} = 0 &\stackrel{\text{plug } \beta_0}{\iff} \sum_{i=1}^n [y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i] x_i = 0 \\ &\iff \sum_{i=1}^n x_i (y_i - \bar{y}) - \beta_1 \sum_{i=1}^n x_i (x_i - \bar{x}) = 0 \\ &\iff \beta_1 = \frac{\sum_{i=1}^n x_i (y_i - \bar{y})}{\sum_{i=1}^n x_i (x_i - \bar{x})} \end{aligned}$$

We can also show that

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

We use a hat on the β 's to show that they are estimates; that is,

$$\begin{aligned} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Call $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ the **fitted values** and $e_i = y_i - \hat{\mu}_i$ the **residual**.

LECTURE 3 | 2020-09-14

Model: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

Equation of fitted line: $y = \hat{\beta}_0 + \hat{\beta}_1 x$

Interpretation:

- $\hat{\beta}_0$ is the estimate of the expected response when $x = 0$ (but not always meaningful if outside range of x_i 's in data)
- $\hat{\beta}_1$ is the estimate of expected change in response for unit increase in x

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x}) \sum_{i=1}^n (y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

- σ^2 is the “variability around the line.”

Recall that $\sigma^2 = \text{Var}[\varepsilon_i] = \text{Var}[Y_i]$

Q: How to estimate σ^2 ?

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_i)$$

$$e_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$$

Intuition: use variability in residuals to estimate σ^2 .

We use

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n - 2}$$

which looks like sample variance of e_i 's. Therefore,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2} = \frac{\text{Ss}[\text{Res}]}{n - 2}$$

Note that “Square Sum” is abbreviated as “Ss”. Now,

$$\bar{e} = \bar{y} - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}) = 0$$

The $n - 2$ will be looked at more carefully later, but for now it suffices to say that $n - 2 = \text{d.f.} = \text{number of parameters estimated}$. It allows $\hat{\sigma}^2$ to be an unbiased estimator for the true value of σ^2 ; that is,

$$\text{E}[\hat{\sigma}^2] = \sigma^2$$

whenever $\hat{\sigma}^2$ is viewed as a random variable.

Q: Is there a statistically significant relationship?

Fact (proved using mgf in STAT 330): Suppose $Y_i \sim N(\mu_i, \sigma_i^2)$ are all independent. Then,

$$\sum_{i=1}^n a_i Y_i \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2 \sigma_i^2\right)$$

for any constant a_i .

In words,

“Linear combination of Normal is Normal.”

Viewing $\hat{\beta}_1$ as a random variable:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y} \sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})x_i - \bar{x} \sum_{i=1}^n (x_i - \bar{x})} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})x_i}$$

So,

$$\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i$$

where $a_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n x_i(x_i - \bar{x})}$.

$$\begin{aligned}
 \mathbb{E}[\hat{\beta}_1] &= \sum_{i=1}^n a_i \mathbb{E}[Y_i] \\
 &= \sum_{i=1}^n a_i (\beta_0 + \beta_1 x_i) \\
 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n x_i(x_i - \bar{x})} \\
 &= \frac{\beta_0 \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} + \beta_1 \sum_{i=1}^n x_i(x_i - \bar{x})}{\sum_{i=1}^n x_i(x_i - \bar{x})} \\
 &= \beta_1
 \end{aligned}$$

On average, $\hat{\beta}_1$ is an unbiased estimator for β_1 .

Now, we calculate the variance of $\hat{\beta}_1$:

$$\begin{aligned}
 \text{Var}[\hat{\beta}_1] &= \sum_{i=1}^n a_i^2 \text{Var}[Y_i] \\
 &= \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n x_i(x_i - \bar{x}) \right]^2} \\
 &= \sigma^2 \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \\
 &= \frac{\sigma^2}{S_{xx}}
 \end{aligned}$$

So, since $\hat{\beta}_1$ is a linear combination of Normals,

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

In a similar manner,

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$$

That is, $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimates.

Then,

$$\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}} \sim N(0, 1)$$

However, σ is unknown, so need to estimate with $\hat{\sigma}$:

$$\frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$$

Since $\text{Sd}[\hat{\beta}_1] = \hat{\sigma}/\sqrt{S_{xx}}$, we say the standard error of $\hat{\beta}_1$ is $\text{Se}[\hat{\beta}_1] = \hat{\sigma}/\sqrt{S_{xx}}$

DEFINITION 0.0.1: Student's T-distribution

T is said to follow a **Student's T-distribution** with k degrees of freedom, denoted $T \sim t(k)$, if

$$T = \frac{Z}{\sqrt{U/k}}$$

where $Z \sim N(0, 1)$ and $U \sim \chi^2(k)$.

Fact: For the simple linear regression model,

$$\frac{\hat{\sigma}^2(n-2)}{\sigma^2} = \frac{\text{Ss}[\text{Res}]}{\sigma^2} \sim \chi^2(n-2)$$

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} = \frac{\frac{\hat{\beta}_1 - \beta_1}{\sigma/\sqrt{S_{xx}}}}{\sqrt{\frac{\hat{\sigma}^2(n-2)}{\sigma^2} \left(\frac{1}{n-2}\right)}} \sim t(n-2)$$

A $(1 - \alpha)$ confidence interval for β_1 is

$$\hat{\beta}_1 \pm (c)\text{Se}[\hat{\beta}_1]$$

where c is the $1 - \frac{\alpha}{2}$ quantile of $t(n-2)$; that is,

- $P(|T| \leq c) = 1 - \alpha$, or
- $P(T \leq c) = 1 - \frac{\alpha}{2}$

where $T \sim t(n-2)$.

Hypothesis test: $H_0: \beta = 0$ versus $H_A: \beta_1 \neq 0$.

If H_0 is true, then

$$\frac{\hat{\beta}_1 - \beta_1}{\text{Se}[\hat{\beta}_1]} = \frac{\hat{\beta}_1}{\text{Se}[\hat{\beta}_1]} \sim t(n-2)$$

so calculate

$$t = \frac{\hat{\beta}_1}{\text{Se}[\hat{\beta}_1]}$$

and reject H_0 at level α if $|t| > c$ where c is $1 - \frac{\alpha}{2}$ quantile of $t(n-2)$.

$$p\text{-value} = P(|T| \geq |t|) = 2P(T \geq |t|)$$

Prediction for SLR: Suppose we want to predict the response y for a new value of x . Say $x = x_0$. Then, SLR model says

$$Y_0 \sim N(\beta_0 + \beta_1 x_0, \sigma^2)$$

where Y_0 is a r.v. for response when $x = x_0$.

The fitted model predicts the *value* of y to be

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

As a random variable,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

then,

$$\mathbb{E}[\hat{Y}_0] = \mathbb{E}[\hat{\beta}_0] + x_0 \mathbb{E}[\hat{\beta}_1] = \beta_0 + \beta_1 x_0 = \mathbb{E}[Y_0]$$

since $\hat{\beta}_i$ for $i = 0, 1$ are unbiased. We can say that \hat{Y}_0 is an unbiased estimate of the random variable for the prediction: Y_0 .

We claim that:

$$\text{Var}[\hat{Y}_0] = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

by expressing $\hat{Y}_0 = \sum_{i=1}^n a_i Y_i$. This implies that,

$$\hat{Y}_0 \sim N \left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

The random variable for prediction error is

$$Y_0 - \hat{Y}_0$$

where Y_0 and \hat{Y}_0 are independent.

$$\mathbb{E}[Y_0 - \hat{Y}_0] = \mathbb{E}[Y_0] - \mathbb{E}[\hat{Y}_0] = 0$$

$$\text{Var}[Y_0 - \hat{Y}_0] = \text{Var}[Y_0] + (-1)^2 \text{Var}[\hat{Y}_0] = \sigma^2 + \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

Again, we have a linear combination of independent Normals, so

$$Y_0 - \hat{Y}_0 \sim N \left(0, \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right) \right)$$

Since σ is unknown, we use $\hat{\sigma}$ and get the following:

$$\frac{Y_0 - \hat{Y}_0}{\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim t(n-2)$$

Intuition for prediction error composed of 2 terms:

- $\text{Var}[Y_0]$: random error of new observation
- $\text{Var}[\hat{Y}_0]$ (predictor): estimating β_0 and β_1

Those are 2 sources of uncertainty.

Note: Be careful that the prediction may not make sense if x_0 is outside the range of the x_i 's in the data.

$(1 - \alpha)$ prediction interval for y_0 :

$$\hat{y}_0 \pm c\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

where c is the $1 - \frac{\alpha}{2}$ quantile of $t(n - 2)$.

Orange production 2018 in FL

- x : acres
- y : # boxes of oranges (thousands)
- (x_i, y_i) recorded for each of 25 FL counties
- $r = 0.964$
- $\bar{x} = 16133$
- $\bar{y} = 1798$
- $S_{xx} = 1.245 \times 10^{10}$
- $S_{xy} = 1.453 \times 10^9$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = 0.1167$$

which is a positive slope (positive correlation between x and y). The expected number of boxes produced is estimated to be about 117 higher per an additional acre.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = -85.3$$

Not meaningful to interpret, since it is the expected production if there were 0 acres (outside the range of x_i) as no county has $x = 0$.

Now suppose

$$Ss[\text{Res}] = 1.31 \times 10^7$$

the residuals are the differences between y_i and the fitted regression line.

- $\hat{\sigma}^2 = \frac{\sum_{i=1}^n e_i^2}{n - 2} = \frac{1.31 \times 10^7}{25 - 2} = 5.7 \times 10^5$
- $Se[\hat{\beta}_1] = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = 0.00676$
- To test $H_0: \beta_1 = 0$, calculate

$$t = \frac{\hat{\beta}_1 - 0}{Se[\hat{\beta}_1]} = \frac{0.1167}{0.00676} \approx 17.3$$

Select the 0.975 quantile (for demonstration purposes) of $t(23)$ is 2.07.

- Note that 17.3 is very unlikely to see in $t(23)$.

Since $17.3 > 2.07$, we reject H_0 at $\alpha = 0.05$ level, conclude there's a significant linear relationship between acres and oranges produced.

The 95% confidence interval for β_1 is

$$0.1167 \pm 2.07(0.00676)$$

which does not contain 0.

$$p\text{-value} = P(|t_{23}| \geq 17.3) = 2P(t_{23} \geq 17.3) \approx 1.2 \times 10^{-14}$$

Predict the # of boxes in thousands produced if we had 10000 acres to grow oranges.

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 = -85.3 + (0.1167)(10000) \approx 1082$$

The 95% prediction interval is:

$$1082 \pm 2.07 \sqrt{5.69 \times 10^5} \sqrt{1 + \frac{1}{25} + \frac{(6133)^2}{1.245 \times 10^{10}}}$$

Note: **not** trying to establish causation.

Check LEARN for `florange.csv`.

Is σ the same for all values of y ?

It appears to be violated, can consider taking the log.

Are the error terms plausibly independent? (e.g. does knowing one e_i help predict e_j for a different county?)

LECTURE 5 | 2020-09-21

0.1 Multiple Linear Regression (MLR)

p explanatory variables which can be categorical, continuous, etc.

Rocket

- x_1 : nozzle area (large or small)
- x_2 : mixture in propellant, ratio oxidized fuel
- Y : thrust

Want to develop linear relationship between y and x_1, x_2, \dots, x_p .

Data n observations each consists of response and explanatory variables $(y_i, x_{i1}, x_{i2}, \dots, x_{ip})$. Then,

$$Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}}_{E[Y_i] = \mu_i}, \sigma^2)$$

or $Y_i = \mu_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$.

We can write in vector/matrix form

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} \beta_0 + \beta_1 x_{11} + \dots + \beta_p x_{1p} \\ \beta_0 + \beta_1 x_{21} + \dots + \beta_p x_{2p} \\ \vdots \\ \beta_0 + \beta_1 x_{n1} + \dots + \beta_p x_{np} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Which we can write as

$$\vec{Y} = X\vec{\beta} + \vec{\varepsilon}$$

where

$$\vec{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}_{n \times 1}, X = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}_{n \times (p+1)}, \vec{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}_{(p+1) \times 1}, \vec{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1}$$

We call $\vec{Y} = (Y_1, Y_2, \dots, Y_n)^\top$ a **random vector** (vector of r.v.'s), analogue of expectation and variance properties.

- Mean vector:

$$\mathbb{E} [\vec{Y}] = \begin{bmatrix} \mathbb{E} [Y_1] \\ \mathbb{E} [Y_2] \\ \vdots \\ \mathbb{E} [Y_n] \end{bmatrix}$$

- Covariance matrix (variance-covariance matrix):

$$\text{Var} [\vec{Y}] = \begin{bmatrix} \text{Var} [Y_1] & \text{Cov} [Y_1, Y_2] & \cdots & \text{Cov} [Y_1, Y_n] \\ \text{Cov} [Y_2, Y_1] & \text{Var} [Y_2] & \cdots & \text{Cov} [Y_2, Y_n] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov} [Y_n, Y_1] & \text{Cov} [Y_n, Y_2] & \cdots & \text{Var} [Y_n] \end{bmatrix}$$

- symmetric since $\text{Cov} [Y_i, Y_j] = \text{Cov} [Y_j, Y_i]$
- positive semi-definite since $\vec{a}^\top \text{Var} [\vec{Y}] \vec{a} \geq 0$ for all $\vec{a} \in \mathbb{R}^n$.
- $\text{Var} [\vec{Y}] = \mathbb{E} \left[(\vec{Y} - \mathbb{E} [\vec{Y}]) (\vec{Y} - \mathbb{E} [\vec{Y}])^\top \right]$

Properties of random vector: let \vec{a} be a $1 \times n$ matrix (row vector) of constants and A be an $n \times n$ matrix of constants.

$$\mathbb{E} [\vec{a}\vec{Y}] = \vec{a}\vec{Y}$$

$$\mathbb{E} [A\vec{Y}] = A\mathbb{E} [\vec{Y}]$$

$$\text{Var} [\vec{a}\vec{Y}] = \vec{a}\text{Var} [\vec{Y}] \vec{a}^\top$$

$$\text{Var} [A\vec{Y}] = A\text{Var} [\vec{Y}] A^\top$$

Derivation of (4):

$$\begin{aligned} \text{Var} [A\vec{Y}] &= \mathbb{E} \left[(A\vec{Y} - \mathbb{E} [A\vec{Y}]) (A\vec{Y} - \mathbb{E} [A\vec{Y}])^\top \right] \\ &= \mathbb{E} \left[(A\vec{Y} - A\mathbb{E} [\vec{Y}]) (A\vec{Y} - A\mathbb{E} [\vec{Y}])^\top \right] \\ &= \mathbb{E} \left[A (\vec{Y} - \mathbb{E} [\vec{Y}]) (A (\vec{Y} - \mathbb{E} [\vec{Y}]))^\top \right] \\ &= \mathbb{E} \left[A (\vec{Y} - \mathbb{E} [\vec{Y}]) (\vec{Y} - \mathbb{E} [\vec{Y}])^\top A^\top \right] \\ &= A\mathbb{E} \left[(\vec{Y} - \mathbb{E} [\vec{Y}]) (\vec{Y} - \mathbb{E} [\vec{Y}])^\top \right] A^\top \\ &= A\text{Var} [\vec{Y}] A^\top \end{aligned}$$

Numerical example: $\vec{Y} = (Y_1, Y_2, Y_3)^\top$. Suppose

$$\mathbb{E} [\vec{Y}] = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

and

$$\text{Var} [Y] = \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix}$$

and

$$\mathbf{a} = [1 \quad -1 \quad 2]$$

and

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Exercise:

- $E[\vec{a}\vec{Y}]$
- $\text{Var}[\vec{a}\vec{Y}]$
- $E[A\vec{Y}]$
- $\text{Var}[A\vec{Y}]$

Let's do the first two,

$$E[\vec{a}\vec{Y}] = \vec{a}E[\vec{Y}] = [1 \quad -1 \quad 2] \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix} = 1(3) - 1(1) + 2(2) = 6$$

$$\begin{aligned} \text{Var}[\vec{a}\vec{Y}] &= [1 \quad -1 \quad 2] \begin{bmatrix} 4 & 1/2 & -2 \\ 1/2 & 1 & 0 \\ -2 & 0 & 3 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \\ &= [1 \quad -1 \quad 2] \begin{bmatrix} 4(1) + (1/2)(-1) - 2(2) \\ (1/2)(1) + 1(-1) + 0(2) \\ -2(1) + 0(-1) + 3(2) \end{bmatrix} \\ &= [1 \quad -1 \quad 2] \begin{bmatrix} -1/2 \\ -1/2 \\ 4 \end{bmatrix} \\ &= 1(-1/2) - 1(-1/2) + 2(4) \\ &= 8 \end{aligned}$$

Multivariate normal distribution (MVN): We say that $\vec{Y} \sim \text{MVN}(\vec{\mu}, \Sigma)$ where $\vec{\mu}$ = mean vector and Σ = covariance matrix. Suppose $Y = (Y_1, \dots, Y_n)^\top$.

$$f(\vec{y}; \vec{\mu}, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\vec{y} - \vec{\mu})^\top \Sigma^{-1} (\vec{y} - \vec{\mu}) \right\}$$

where Σ^{-1} is the inverse of the covariance matrix and $|\Sigma|$ is the determinant of Σ .

Properties of MVN: Suppose $\vec{Y} = (Y_1, \dots, Y_n)^\top \sim \text{MVN}(\vec{\mu}, \Sigma)$ and \vec{a} is a $1 \times n$ constant and A is an $n \times n$ matrix of constants.

1. Linear transformations of MVN is MVN, so

$$\vec{a}\vec{Y} \sim \text{MVN}(\vec{a}\vec{\mu}, \vec{a}\Sigma\vec{a}^\top)$$

$$A\vec{Y} \sim \text{MVN}(A\vec{\mu}, A\Sigma A^\top)$$

2. Marginal distribution of Y_i is Normal,

$$Y_i \sim N(\mu_i, \Sigma_{ii})$$

In fact, any subset of Y_i 's is MVN

3. Conditional MVN is MVN, e.g. $Y_1 \mid Y_2, \dots, Y_n$

4. Another property:

$$\text{Cov}[Y_i, Y_j] = 0 \iff Y_i, Y_j \text{ independent}$$

that is, Y_i and Y_j are uncorrelated.

$$\Sigma_{ij} = 0$$