Chapter 5. Vector random variables

- A **vector random variable** $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ is a collection of random numbers with probabilities assigned to outcomes.
- $\mathbf{X}$ can also be called a **multivariate random variable**.
- The case with $n = 2$ we call a **bivariate random variable**.
- Saying X and Y are **jointly distributed random variables** is equivalent to saying (X, Y) is a bivariate random variable.
- Vector random variables let us model relationships between quantities.

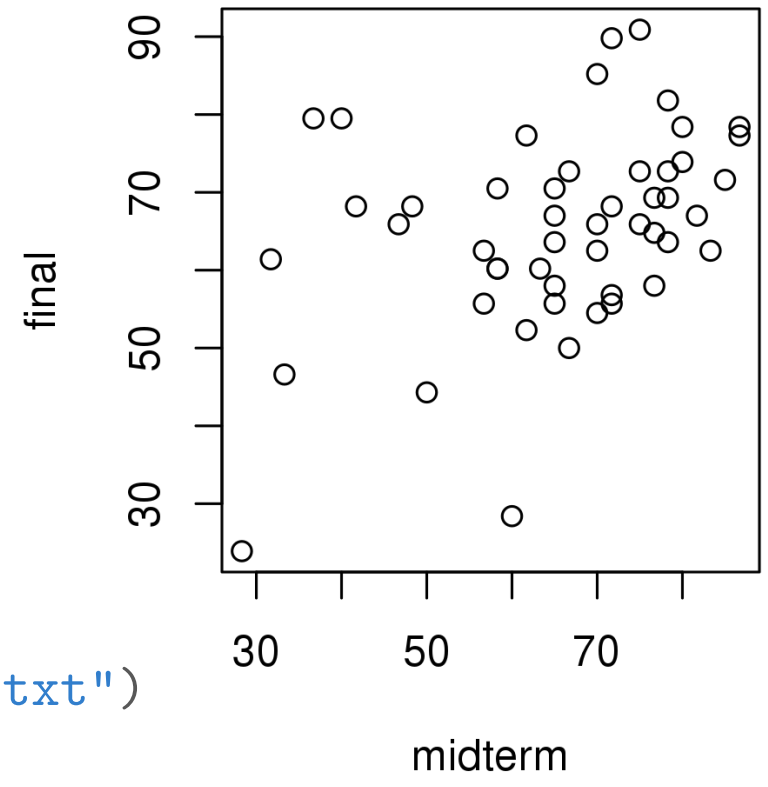
Example: midterm and final scores

- We will look at the anonymized test scores for a previous course.

```
download.file(destfile="course_progress.txt",
 url="https://ionides.github.io/401f18/05/course_progress.txt")
```

```
# Anonymized scores for a random subset of 50 students
"final" "quiz" "hw" "midterm"
"1" 52.3 76.7 91 61.7
"2" 68.2 65.4 94.5 48.3
"3" 78.4 91.2 95.5 80
```

- A probability model lets us answer a question like, "What is the probability that someone gets at least 70% in both the midterm and the final"




```
x <- read.table("course_progress.txt")
plot(final~midterm,data=x)
```

The bivariate normal distribution and covariance

- Let $X \sim \text{normal}(\mu_X, \sigma_X)$ and $Y \sim \text{normal}(\mu_Y, \sigma_Y)$.
- If X and Y are bivariate random variables we need another parameter to describe their dependence. If X is big, does Y tend to be big, or small, or does the value of X make no difference to the outcome of Y?
- This parameter is the **covariance**, defined to be

$$\text{Cov}(X, Y) = \text{E}\Big[\big(X - \text{E}[X]\big)\big(Y - \text{E}[Y]\big)\Big]$$

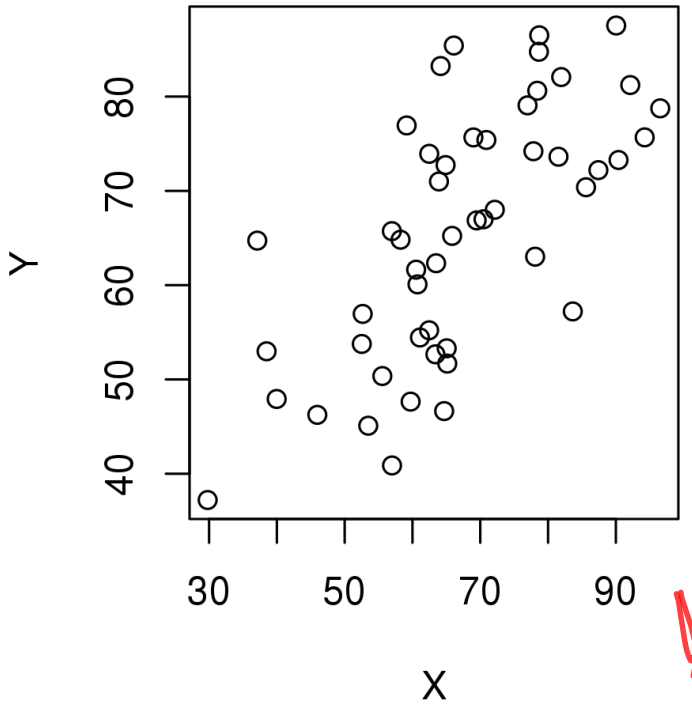
- The parameters of the bivariate normal distribution in matrix form are the **mean vector** $\boldsymbol{\mu} = (\mu_X, \mu_Y)$ and the **variance/covariance matrix**,

$$\mathbb{V} = \begin{bmatrix} \text{Var}(X) & \text{Cov}(X, Y) \\ \text{Cov}(Y, X) & \text{Var}(Y) \end{bmatrix}$$

install.packages("mvtnorm") ; library("mvtnorm")

- In R, the `mvtnorm` package lets us simulate the bivariate and multivariate normal distribution via the `rmvnorm()` function. It has the mean vector and variance/covariance matrix as arguments.

Experimenting with the bivariate normal distribution




```
library(mvtnorm)
mvn <- rmvnorm(n=50,
  mean=c(X=65,Y=65),
  sigma=matrix(
    c(200,100,100,150),
    2,2)
)
plot(Y~X,data=mvn)
```

$\mathbb{V} = \begin{bmatrix} 200 & 100 \\ 100 & 150 \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Var}(Y) \end{bmatrix}$

- We write $(X, Y) \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbb{V})$, where MVN is read "multivariate normal".

Question 5.1. What are μ_X, μ_Y, $\mathrm{Var}(X)$, $\mathrm{Var}(Y)$, and $\mathrm{Cov}(X, Y)$ for this simulation?

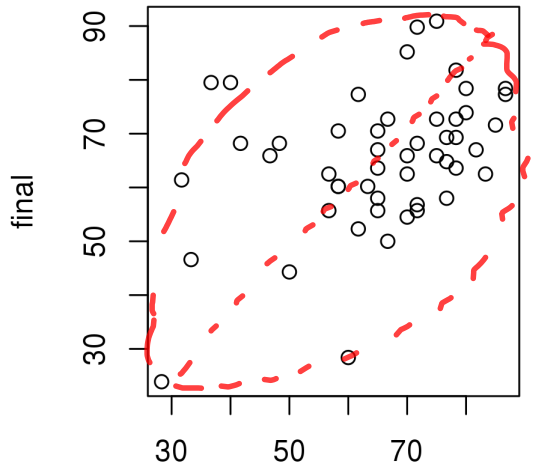
$\mu_X = 65$, $\mu_Y = 65$, $\mathrm{Var}(X) = 200$, $\mathrm{Var}(Y) = 150$

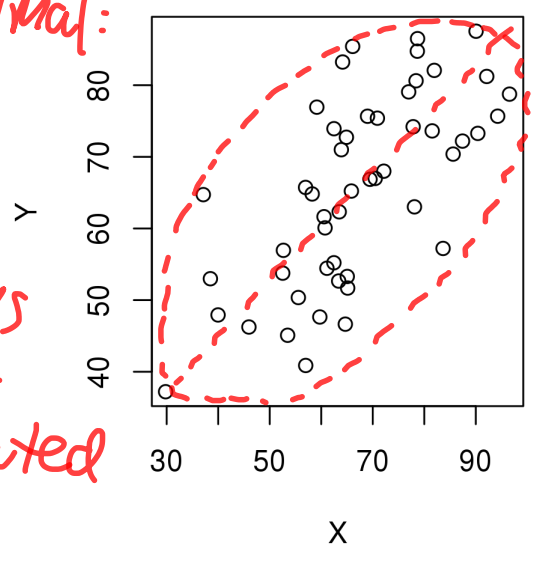
$\mathrm{Cov}(X,Y) = 100$

The bivariate normal as a model for exam scores

Question 5.2. Compare the data on midterm and final scores with the simulation. Does a normal model seem to fit? Would you expect it to? Why, and why not?

- Maybe the exam scores are clustered more in the top right?
- Maybe the exam scores are more variable; perhaps the distribution is longer tailed than normal?
- So far as the scatterplot does look normal: both are scattered around a line in something like a football shape.
- A central limit theorem here: both exams are the sum of many questions. The test-taking skill may not be normally distributed but the test uncertainty should be.





More on covariance

- Covariance is **symmetric**: we see from the definition

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathrm{E}\Big[\big(X-\mathrm{E}[X]\big)\big(Y-\mathrm{E}[Y]\big)\Big] \\ &= \mathrm{E}\Big[\big(Y-\mathrm{E}[Y]\big)\big(X-\mathrm{E}[X]\big)\Big] \;=\; \mathrm{Cov}(Y,X)\end{aligned}$$

- Also, we see from the definition that $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$.
- The **sample covariance** of n pairs of measurements $(x_1,y_1),\ldots,(x_n,y_n)$ is

$$\mathrm{cov}(\mathbf{x},\mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

where $\bar{x}$ and $\bar{y}$ are the sample means of $\mathbf{x}=(x_1,\ldots,x_n)$ and $\mathbf{y}=(y_1,\ldots,y_n)$.

Scaling covariance to give correlation

- The standard deviation of a random variable is interpretable as its scale.
- Variance is interpretable as the square of standard deviation

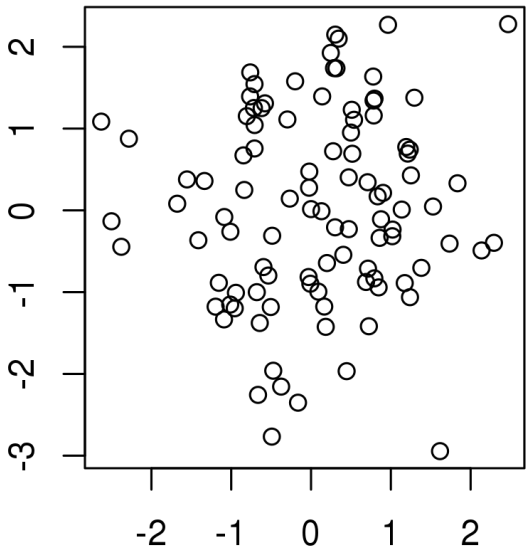
```
var(x$midterm)
## [1] 218.2155
var(x$final)
## [1] 169.7518
cov(x$midterm,x$final)
## [1] 75.61269
```

- Covariance is interpretable when scaled to give the **correlation**

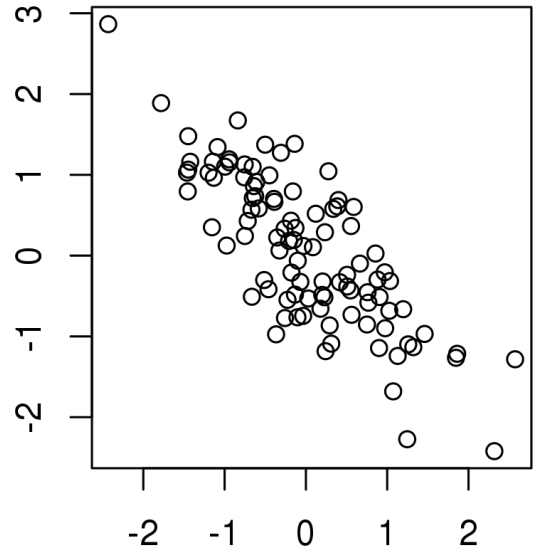
$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \approx \frac{\mathrm{Cov}(X,Y)}{SD(X)\,SD(Y)} \mathrm{cor}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}}$$

```
cor(x$midterm,x$final)
## [1] 0.3928662
```

```
rho <- 0
```

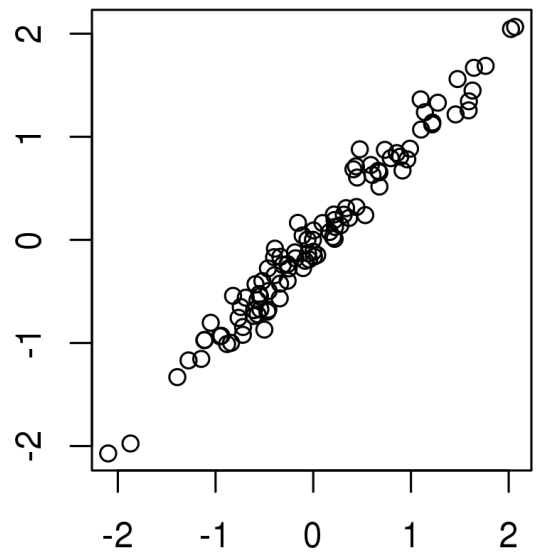



```
rho <- -0.8
```




```
library(mvtnorm)
mvn <- rmvnorm(n=100,
  mean=c(X=0,Y=0),
  sigma=matrix(
    c(1,rho,rho,1),
    2,2)
)
```

```
rho <- 0.95
```



More on interpreting correlation

- Random variables with a correlation of ± 1 (or data with a sample correlation of ± 1) are called **linearly dependent** or **colinear**.
- Random variables with a correlation of 0 (or data with a sample correlation of 0) are **uncorrelated**.
- Random variables with a covariance of 0 are also uncorrelated!

Question 5.3. Suppose two data vectors $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ have been **standardized**. That is, each data point has had the sample mean substracted and then been divided by the sample standard deviation. You calculate $\text{cov}(\mathbf{x}, \mathbf{y}) = 0.8$. What is the sample correlation, $\text{cor}(\mathbf{x}, \mathbf{y})$?

$$\text{cor}(\underset{\sim}{x}, \underset{\sim}{y}) = \frac{\text{cov}(\underset{\sim}{x}, \underset{\sim}{y})}{\underbrace{\text{sd}(\underset{\sim}{x})\,\text{sd}(\underset{\sim}{y})}_{1\times 1}} = \text{cov}(\underset{\sim}{x}, \underset{\sim}{y})$$

The variance of a sum

- A basic property of covariance is

(Eq. C1) $$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$$

- Sample covariance has the same formula,

(Eq. C2) $$\mathrm{var}(\mathbf{x}+\mathbf{y}) = \mathrm{var}(\mathbf{x}) + \mathrm{var}(\mathbf{y}) + 2\,\mathrm{cov}(\mathbf{x},\mathbf{y})$$

- These nice formulas mean it can be easier to calculate using variances and covariances rather than standard deviations and correlations.

Question 5.4. Rewrite (Eq. C1) to give a formula for $\mathrm{SD}(X+Y)$ in terms of $\mathrm{SD}(X)$, $\mathrm{SD}(Y)$ and $\mathrm{Cor}(X,Y)$.

$$[SD(X+Y)]^2 = (SD(X))^2 + (SD(Y))^2 + 2\,Cor(X,Y)\,SD(X)\,SD(Y)$$

So, $$SD(X+Y) = \sqrt{(SD(X))^2 + (SD(Y))^2 + 2\,Cor(X,Y)\,SD(X)\,SD(Y)}.$$

More properties of covariance

* ie. for (X, Y, Z) a vector random variable.

- Covariance is not affected by adding constants to either variable

(Eq. C3) $$\mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y)$$

- Recall the definition $\mathrm{Cov}(X, Y) = \mathrm{E}\big[\big(X - \mathrm{E}[X]\big)\big(Y - \mathrm{E}[Y]\big)\big]$. In words, covariance is the mean product of deviations from average. These deviations are unchanged when we add a constant to the variable.
- Covariance scales **bilinearly** with each variable

(Eq. C3) $$\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y)$$

- Covariance distributes across sums * for X, Y, Z joint random variables

(Eq. C4) $$\mathrm{Cov}(X, Y + Z) = \mathrm{Cov}(X, Y) + \mathrm{Cov}(X, Z)$$

- Sample covariances also have these properties. You can test them in R using bivariate normal random variables, constructed as previously using 'rmvnorm()'.

e.g. $\underset{\sim}{x} = (x_1, \ldots, x_n)$, $\underset{\sim}{y} = (y_1, \ldots, y_n)$, $\underset{\sim}{z} = (z_1, \ldots, z_n)$ three vectors with the same length n.

The variance/covariance matrix of vector random variables

- Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector random variable. For any pair of elements, say X_i and X_j, we can compute the usual scalar covariance, $v_{ij} = \mathrm{Cov}(X_i, X_j)$.
- The variance/covariance matrix $\mathbb{V} = [v_{ij}]_{p \times p}$ collects together all these covariances.

$$\mathbb{V} = \mathrm{Var}(\mathbf{X}) = \begin{bmatrix} \mathrm{Cov}(X_1, X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Cov}(X_2, X_2) & & \mathrm{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Cov}(X_p, X_p) \end{bmatrix}$$

- The diagonal entries of $\mathbb{V}$ are $v_{ii} = \mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$ for $i = 1, \ldots, p$ so the variance/covariance matrix can be written as

$$\mathbb{V} = \mathrm{Var}(\mathbf{X}) = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \ldots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & & \mathrm{Cov}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cov}(X_p, X_1) & \mathrm{Cov}(X_p, X_2) & \ldots & \mathrm{Var}(X_p) \end{bmatrix}$$

The correlation matrix

- Covariance is harder to interpret than correlation, but easier for calculations.
- We can put together all the correlations into a correlation matrix, using the fact that $\mathrm{Cor}(X_i, X_i) = 1$.

$$\mathrm{Cor}(\mathbf{X}) = \begin{bmatrix} 1 & \mathrm{Cor}(X_1, X_2) & \dots & \mathrm{Cor}(X_1, X_p) \\ \mathrm{Cor}(X_2, X_1) & 1 & & \mathrm{Cor}(X_2, X_p) \\ \vdots & & \ddots & \vdots \\ \mathrm{Cor}(X_p, X_1) & \mathrm{Cor}(X_p, X_2) & \dots & 1 \end{bmatrix}$$

- Multivariate distributions can be very complicated.
- The variance/covariance and correlation matrices deal only with **pairwise** relationships between variables.
- Pairwise relationships can be graphed.

The sample variance/covariance matrix

- The **sample variance/covariance matrix** places all the sample variances and covariances in a matrix.
- Let $\mathbb{X} = [x_{ij}]_{n \times p}$ be a data matrix made up of p data vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ each of length n. (check the case p=2 for intuition)

(0\0) not (0/0)

$$\underbrace{\text{var}(\mathbb{X})}_{p \times p} = \begin{bmatrix} \text{var}(\mathbf{x}_1) & \text{cov}(\mathbf{x}_1, \mathbf{x}_2) & \ldots & \text{cov}(\mathbf{x}_1, \mathbf{x}_p) \\ \text{cov}(\mathbf{x}_2, \mathbf{x}_1) & \text{var}(\mathbf{x}_2) & & \text{cov}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & & \ddots & \vdots \\ \text{cov}(\mathbf{x}_p, \mathbf{x}_1) & \text{cov}(\mathbf{x}_p, \mathbf{x}_2) & \ldots & \text{var}(\mathbf{x}_p) \end{bmatrix}$$

- R uses the same notation. If `x` is a matrix or dataframe, `var(x)` returns the sample variance/covariance matrix.

Note: the diagonal elements of a matrix are the (1,1), (2,2), ..., (p,p) entries which give $\text{Var}(x_1), \ldots \text{Var}(x_p)$.

```
var(x)
##               final      quiz         hw   midterm
## final    169.75184  78.14294  51.27143  75.61269
## quiz      78.14294 224.39664 103.57755 107.32550
## hw        51.27143 103.57755 120.13265  61.44694
## midterm   75.61269 107.32550  61.44694 218.21553
```

The sample correlation matrix

- The **sample correlation matrix** places all the sample correlations in a matrix.
- Let $\mathbb{X} = [x_{ij}]_{n \times p}$ be a data matrix made up of p data vectors $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_p}$ each of length n.

$$\text{cor}(\mathbb{X}) = \begin{bmatrix} 1 & \text{cor}(\mathbf{x}_1, \mathbf{x}_2) & \ldots & \text{cor}(\mathbf{x}_1, \mathbf{x}_p) \\ \text{cor}(\mathbf{x}_2, \mathbf{x}_1) & 1 & & \text{cor}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & & \ddots & \vdots \\ \text{cor}(\mathbf{x}_p, \mathbf{x}_1) & \text{cor}(\mathbf{x}_p, \mathbf{x}_2) & \ldots & 1 \end{bmatrix}$$

- R uses the same notation. If `x` is a matrix or dataframe, `cor(x)` returns the sample correlation matrix.

```
cor(x)
```

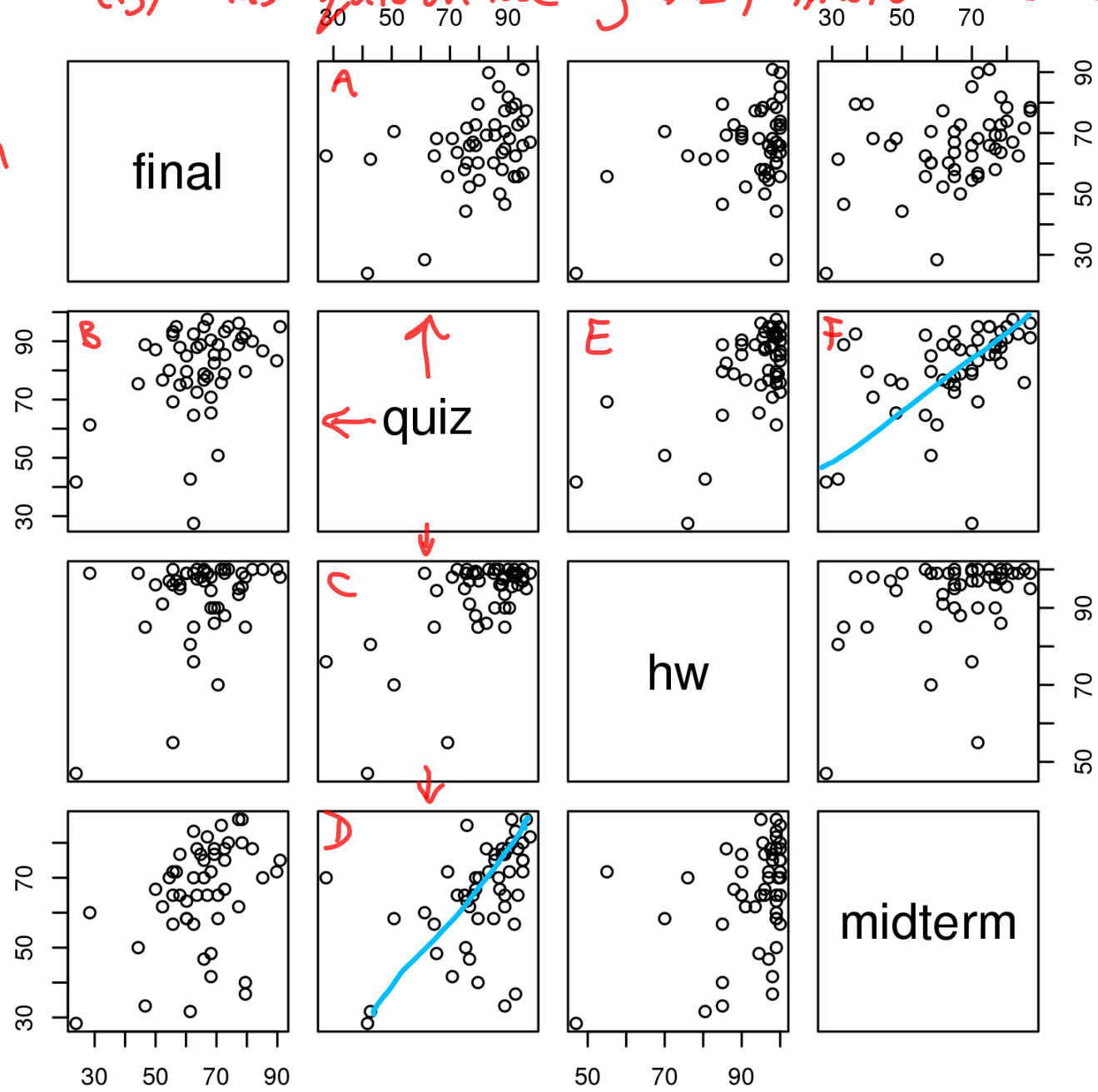
```
##               final      quiz        hw   midterm
## final    1.0000000 0.4003818 0.3590357 0.3928662
## quiz     0.4003818 1.0000000 0.6308512 0.4850114
## hw       0.3590357 0.6308512 1.0000000 0.3795132
## midterm 0.3928662 0.4850114 0.3795132 1.0000000
```

```
pairs(x)
```

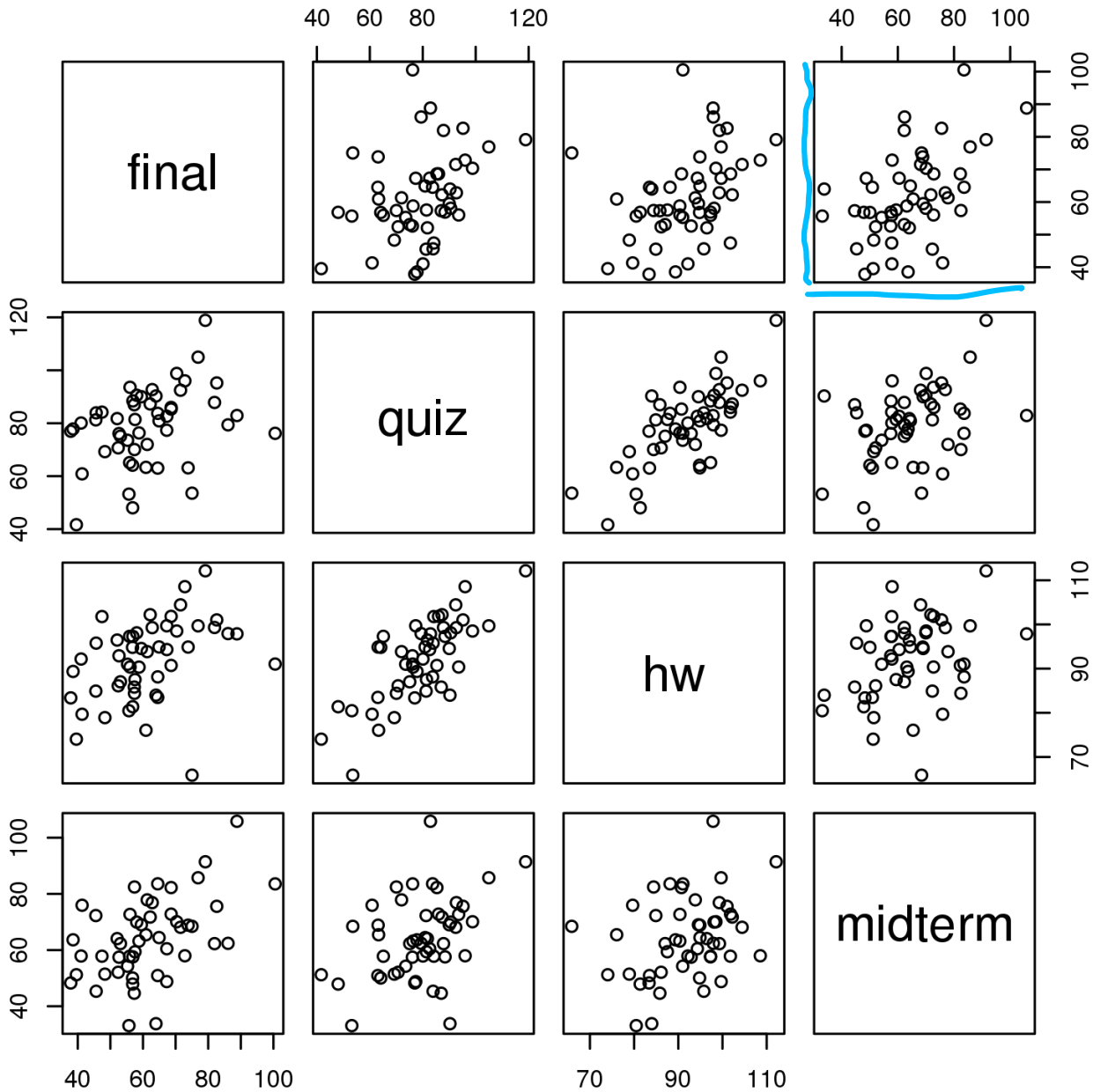
(A) has quiz on the x axis, final on the y axis.
(B) has quiz on the y axis, final on the x axis.

(A,C,D) all have quiz on the x-axis

(B,E,F) all have quiz on the y axis



```
mvn <- rmvnorm(50,mean=apply(x,2,mean),sigma=var(x))
pairs(mvn)
```



Question 5.5. From looking at the scatterplots, what are the strengths and weaknesses of a multivariate normal model for test scores in this course? Specifically, why does HW not follow a normal curve?

Difficulty of the assignment? Office hrs?
Maybe HW tests attention to detail?
This means the central limit theorem is not holding.
Why not? Questions not independent - all too similar?
Each HW is all or nothing, so there are not a large number.
HW score is driven by a small number of rare events – missed HW.

Question 5.6. To what extent is it appropriate to summarize the data by the mean and variance/covariance matrix (or correlation matrix) when the normal approximation is dubious?

multivariate
The normal distribution is fully determined by the mean & variance/covariance matrix. A football-shaped scatterplot is well described by the sample mean & sample variance/covariance matrix. If the scatterplot is not football shaped, there is information lost by summarizing it as mean & variance/covariance.

Linear combinations

here, $\underset{\sim}{x} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}$ is a column vector, written this way to match $\underset{\sim}{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$

- Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector random variable with mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ and $p \times p$ variance/covariance matrix $\mathbb{V}$.
- Let $\mathbb{X}$ be a $n \times p$ data matrix.
- Let $\mathbb{A}$ be a $q \times p$ matrix.
- $\mathbf{Z} = \mathbb{A}\mathbf{X}$ is a collection of q linear combinations of the p random variables in the vector $\mathbf{X}$, viewed as a **column** vector.
- $\mathbb{Z} = \mathbb{X}\mathbb{A}^{\mathrm{T}}$ is an $n \times q$ collection of linear combinations of the p data points in each **row** of $\mathbb{X}$.
- Mental gymnastics are required: vectors are often interpreted as **column vectors** (e.g., $p \times 1$ matrices) but the vector of measurements for each unit is a **row vector** when considered as a row of an $n \times p$ data matrix.

we are going to be interested in $(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\underset{\sim}{y}$ for a probability model for the linear model.

$\mathbf{Z} = \mathbb{A}\mathbf{X}$: $q \times 1$ $q \times p$ $p \times 1$

$\mathbb{Z} = \mathbb{X}\mathbb{A}^{\mathrm{T}}$: $n \times q$

this will be a fixed matrix $\mathbb{A}$

this will be our random variable $\underset{\sim}{X}$

one row of $\mathbb{Z}$ is $\underbrace{[x_{i1}\ x_{i2} \ldots x_{ip}]}_{\underset{\sim}{x}^T}\mathbb{A}^T = \underset{\sim}{x}^T\mathbb{A}^T$

transposing this gives a column vector $(\underset{\sim}{x}^T\mathbb{A}^T)^T = \mathbb{A}\underset{\sim}{x}$.

Variables of length p as column vectors or row vectors

Question 5.7. How would you construct a simulated data matrix $\mathbb{Z}_{\text{sim}}$ from n realizations $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ of the random column vector $\mathbf{Z} = \mathbb{A}\mathbf{X}$? Hint: You are expected to write a matrix constructing $\mathbb{Z}_{\text{sim}}$ by stacking together $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$. Be careful with transposes and keep track of dimensions. Recall that $\mathbb{Z}_{\text{sim}}$ should be $n \times p$.

$$\mathbb{Z}_{sim} = \begin{bmatrix} Z_1^T \\ Z_2^T \\ \vdots \\ Z_n^T \end{bmatrix}$$

Expectation and variance of linear combinations

- The expectation of a vector random variable is the vector of expectations of each element. If $\mathbf{X} = (X_1, \ldots, X_n)$ then

$$\mathrm{E}[\mathbf{X}] = \big(\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_n]\big)$$

intuition: expectation is a long run average. averages have this additive property.

- The expectation of a sum is the sum of the expectations.

$$\boxed{\mathrm{E}[X + Y] = \mathrm{E}[X] + \mathrm{E}[Y]}$$

- This formula extends to any linear combination of n random variables. If $\mathbf{Z} = \mathbb{A}\mathbf{X}$ then $\mathrm{E}[\mathbf{Z}] = \mathbb{A}\,\mathrm{E}[\mathbf{X}]$. In other words,

$$\boxed{\mathrm{E}[\mathbb{A}\,\mathbf{X}] = \mathbb{A}\,\mathrm{E}[\mathbf{X}]}$$

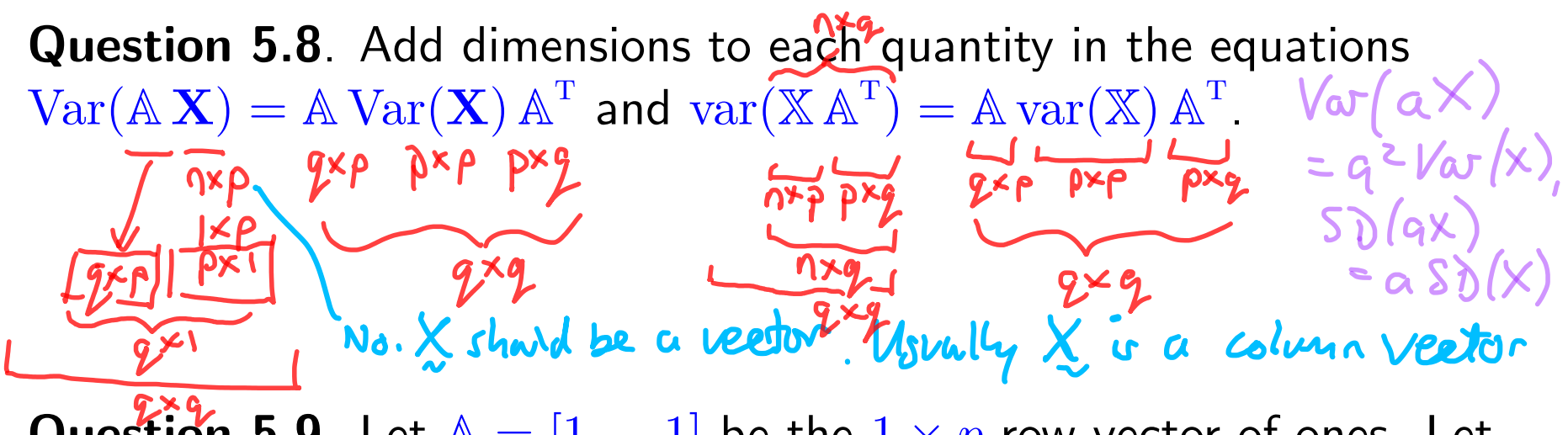
- There is a useful matrix variance/covariance formula for a linear combination, which also works for sample variance/covariance.

$$\boxed{\mathrm{Var}(\mathbb{A}\,\mathbf{X}) = \mathbb{A}\,\mathrm{Var}(\mathbf{X})\,\mathbb{A}^{\mathrm{T}}}$$

$$\boxed{\mathrm{var}(\mathbb{X}\,\mathbb{A}^{\mathrm{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathrm{T}}}$$

Exercises with the matrix variance/covariance formula

Question 5.8. Add dimensions to each quantity in the equations $\mathrm{Var}(\mathbb{A}\,\mathbf{X}) = \mathbb{A}\,\mathrm{Var}(\mathbf{X})\,\mathbb{A}^{\mathrm{T}}$ and $\mathrm{var}(\mathbb{X}\,\mathbb{A}^{\mathrm{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathrm{T}}$.



$\mathrm{Var}(aX) = a^2\mathrm{Var}(X)$, $\mathrm{SD}(aX) = a\,\mathrm{SD}(X)$

Question 5.9. Let $\mathbb{A} = [1 \ldots 1]$ be the $1 \times p$ row vector of ones. Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a vector random variable (treated as a column vector) with variance/covariance matrix $\mathbb{V} = [V_{ij}]_{p\times p}$. Evaluate the variance/covariance formula in this case. Hence, find $\mathrm{Var}(\bar{X})$ where $\bar{X} = (1/p)\sum_{i=1}^{p} X_i$.

What is $\mathbb{A}\underset{\sim}{X}$? $\mathbb{A}\underset{\sim}{X} = (1\ 1 \ldots 1)\begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} = \sum_{i=1}^{p} X_i$

this multiplication sums over rows in each column

this multiplication sums columns in each row

What is $\mathrm{Var}(\mathbb{A}\underset{\sim}{X})$? Using the variance/covariance formula,

$$\mathrm{Var}(\mathbb{A}\underset{\sim}{X}) = (1\ 1 \ldots 1)\begin{pmatrix} V_{11} & V_{12} & \ldots & V_{1p} \\ \vdots & & & \\ V_{p1} & \ldots & & V_{pp} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \sum_{i=1}^{p}\sum_{j=1}^{p} V_{ij}$$

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{p}\mathbb{A}\underset{\sim}{X}\right) = \frac{1}{p^2}\mathrm{Var}(\mathbb{A}\underset{\sim}{X}) = \frac{1}{p^2}\sum_{i=1}^{p}\sum_{j=1}^{p} V_{ij}$$

Testing the variance/covariance formula

- Suppose that the overall course score is weighted 40% on the final and 20% on each of the miterm, homework and quiz.
- We can find the sample variance of the overall score two different ways.

(i) Directly computing the overall score for each student.

as.matrix turns the dataframe x into a matrix.

```
weights <- c(final=0.4,quiz=0.2,hw=0.2,midterm=0.2)
overall <- as.matrix(x) %*% weights
var(overall)
```

50×4 · 4×1 → 50×1

```
##             [,1]
## [1,] 104.2624
```

(ii) Using $\mathrm{var}(\mathbb{X}\,\mathbb{A}^{\mathrm{T}}) = \mathbb{A}\,\mathrm{var}(\mathbb{X})\,\mathbb{A}^{\mathrm{T}}$.

1×4 · 4×4 · 4×1

```
weights %*% var(x) %*% weights
```

1×1

```
##             [,1]
## [1,] 104.2624
```

- R interprets the vector 'weights' as a row or column vector as necessary.

Independence

- Two events E and F are **independent** if

$$\mathrm{P}(E \text{ and } F) = \mathrm{P}(E) \times \mathrm{P}(F)$$

Worked example 5.1. Suppose we have a red die and a blue die. They are idea fair dice, so the values should be independent. What is the chance they both show a six?

(a) Using the definition of independence.

P({red die shows 6} and {blue die shows 6})
= P(red die shows 6) × P(blue die shows 6) : independence
= $\frac{1}{6} \times \frac{1}{6}$ since each die is fair = $\frac{1}{36}$

(b) By considering equally likely outcomes, without using the definition.

There are 36 possible pairs of outcomes (red die, blue die). Assuming equally likely outcomes, each pair (red, blue) has probability 1/36, so P(both dice show 6) = $\frac{1}{36}$.

- The multiplication rule agrees with an intuitive idea of independence.

Independence of random variables

- X and Y are **independent random variables** if, for any intervals $[a, b]$ and $[c, d]$,

$$\mathrm{P}(a < X < b \text{ and } c < Y < d) = \mathrm{P}(a < X < b) \times \mathrm{P}(c < Y < d)$$

- This definition extends to vector random variables. $\mathbf{X} = (X_1, \ldots, X_n)$ is a **vector of independent random variables** if for any collection of intervals $[a_i, b_i]$, $1 \leq i \leq n$,

$$\mathrm{P}(a_1 < X_1 < b_1, \ldots, a_n < X_n < b_n) = \mathrm{P}(a_1 < X_1 < b_1) \times \cdots \times \mathrm{P}(a_n < X_n < b_n)$$

- $\mathbf{X} = (X_1, \ldots, X_n)$ is a **vector of independent identically distributed (iid) random variables** if, in addition, each element of $\mathbf{X}$ has the same distribution.
- "$X_1, \ldots, X_n$ are n random variables with the $\mathrm{normal}(\mu, \sigma)$ distribution" is written more formally as "Let $X_1, \ldots, X_n \sim \text{iid } \mathrm{normal}(\mu, \sigma)$."

Independent vs uncorrelated

a property of random variables, not data

a property of random variables, with a comparable sample version.

- If X and Y are independent they are uncorrelated.
- The converse is not necessarily true.
- **For normal random variables, the converse is true**.
- If X and Y are bivariate normal random variables, and $\mathrm{Cov}(X, Y) = 0$, then X and Y are independent.
- The following slide demonstrated the possibility of being uncorrelated but not independent (for non-normal random variables).
- If the scatter plot of two variables looks normal and their sample correlation is small, the variables are appropriately modeled as independent.

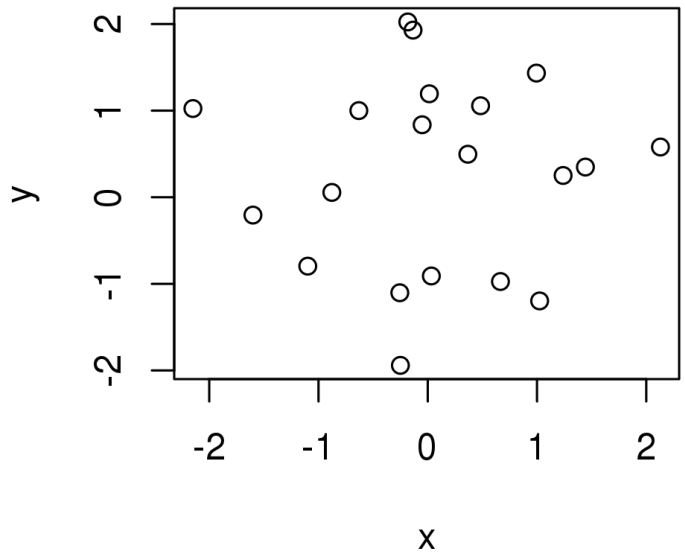
Zero correlation with and without independence

```r
x <- rnorm(20)
y <- rnorm(20)
```

```r
cor(x,y)

## [1] 0.01825057
```

```r
plot(x,y)
```

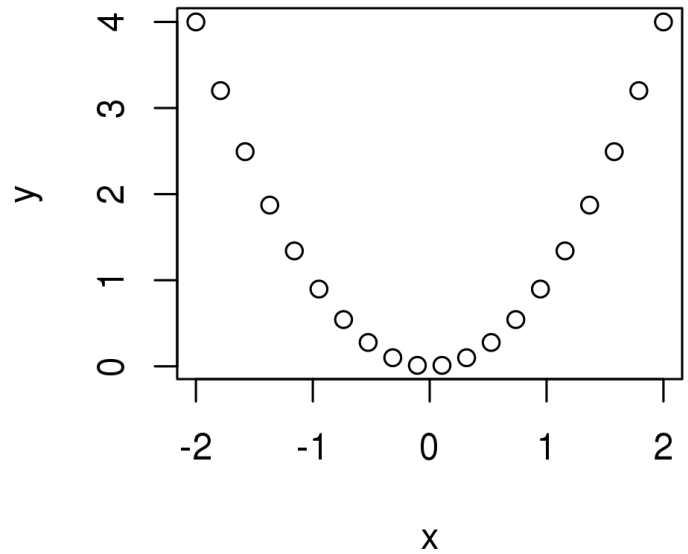



```r
x <- seq(-2,2,length=20)
y <- x^2
```

```r
cor(x,y)

## [1] -1.704156e-16
```

```r
plot(x,y)
```



The measurement error model

- Let $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$ be a vector consisting of n independent normal random variables, each with mean zero and variance σ^2.
- This is a common model for **measurement error** on n measurements.
- The mean vector and variance/covariance matrix are

$$\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}, \qquad \mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2 \mathbb{I}$$

where $\mathbf{0} = (0, \ldots, 0)$ and $\mathbb{I}$ is the $n \times n$ identity matrix.

- For the measurement error model, we break our usual rule of using upper case letters for random variables.
- We can write $\boldsymbol{\epsilon} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2 \mathbb{I})$ or $\boldsymbol{\epsilon} \sim \text{iid normal}(0, \sigma)$.
- $\boldsymbol{\epsilon}$ models **unbiased** measurements (meaning $\mathrm{E}[\epsilon_i] = 0$) subject to **additive Gaussian error**.

Example: 5 repeated measurements $x_1, \ldots, x_5$ of the speed of light could be modeled as $X_i = \mu + \epsilon_i$ for $i = 1, \ldots, 5$, where μ is the unknown true value of this quantity.

A probability model for the linear model

- First recall the sample version of the linear model, which is

$$\mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e},$$

Constant, not random.

vector random variable

where $\mathbf{y}$ is the measured response, $\mathbb{X}$ is an $n \times p$ matrix of explanatory variables, $\mathbf{b}$ is chosen by least squares, and $\mathbf{e}$ is the vector of residuals.

- We want to build a random vector $\mathbf{Y}$ that provides a probability model for the data $\mathbf{y}$. We write this as

$$\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

vector random variable

A probability model in principle defines the probability of everything in the model.

where $\mathbb{X}$ is the explanatory matrix, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is an unknown coefficient vector (we don't know the true population coefficient!) and $\boldsymbol{\epsilon}$ is Gaussian measurement error with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$.

- Our probability model asserts that the process which generated the response data $\mathbf{y}$ was like drawing a random vector $\mathbf{Y}$ constructed using a random measurement error model with known matrix $\mathbb{X}$ for some fixed but unknown value of $\boldsymbol{\beta}$.

we treat $\mathbb{X}$ as a constant matrix.

A digression on "useful" models

"Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an *ideal* gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules. For such a model there is no need to ask the question 'Is the model true?'. If *truth* is to be the *whole truth* the answer must be *No*. The only question of interest is 'Is the model illuminating and useful.' " (Box, 1978)

"**Essentially, all models are wrong, but some are useful.**" (Box and Draper, 1987)

- Perhaps the most useful statistical model ever is $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- Anything so widely used is also widely abused. Our task is to understand $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ so that we can use it wisely.

Expectation and variance/covariance of $\mathbf{Y}$

or $E[\mathbb{X}\beta + \epsilon] = E[\mathbb{X}\beta] + E[\epsilon] = \mathbb{X}\beta + 0 = \mathbb{X}\beta$ $\quad \beta = (\beta_1, \ldots, \beta_p)$

- Recall the linear model $\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbb{X}}\underset{p\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}}$ where $\boldsymbol{\epsilon} \sim \text{MVN}(\mathbf{0}, \sigma^2\mathbb{I})$.

Question 5.10. What is the expected value, $\mathrm{E}[\mathbf{Y}]$?

$E[Y]$ is a constant vector of length n. The i^{th} element of $E[Y]$ is $E[\sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i]$

since $Y_i = \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\mathbb{X} = [x_{ij}]$

so, $E[Y_i] = \sum_{j=1}^{p} x_{ij}\beta_j + E[\epsilon_i]$ $\quad E[\epsilon_i] = 0$

since expectation of a sum is the sum of expectations, and expectation of a constant is constant

so, $E[Y_i] = \sum_{j=1}^{p} x_{ij}\beta_j$

in matrix notation, $E[Y] = \mathbb{X}\beta$

Question 5.11. What is the variance/covariance matrix, $\mathrm{Var}(\mathbf{Y})$?

$$\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{X}\beta + \epsilon) = \mathrm{Var}(\epsilon) = \sigma^2\mathbb{I}$$

this is a constant vector, by assumption, and adding a constant doesn't change variance

by assumption of the measurement error model.

[Note: assuming $\mathbb{X}$ is constant is like conditioning on $\mathbb{X}$ if $\mathbb{X}$ is random — but we don't pursue this idea]

Expectation of the least squares coefficient

this is a random vector

Worked example 5.2. Let $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$. This is the probability model for the sample least squares coefficient $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$. Use linearity to calculate $\mathrm{E}[\hat{\boldsymbol{\beta}}]$.

Expectation is linear means that it preserves linear relationships. if $Y = aX + b$, $E[Y] = a\,E[X] + b$

Solution:

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{\beta}}] &= \mathrm{E}\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}\right] \\
&= \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\,\mathrm{E}[\mathbf{Y}] \\
&= \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbb{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}$$

from previous slide

- Interpretation: If the data $\mathbf{y}$ are well modeled as a draw from the probability model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the least squares estimate $\mathbf{b}$ is well modeled by a random vector with mean $\boldsymbol{\beta}$.

Variance/covariance matrix of the least squares coefficients

Question 5.12. Consider the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$. Use the formula $\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^{\mathrm{T}}$ to find $\mathrm{Var}(\hat{\boldsymbol{\beta}})$.
Hint: what should be our choice of $\mathbb{A}$ so that $\hat{\boldsymbol{\beta}} = \mathbb{A}\mathbf{Y}$?

scalar the constant σ^2 comes out of the product. The Matrix $\mathbb{I}$ disappears since $\mathbb{I}\mathbb{X} = \mathbb{X}$

$$(\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbb{X})(\mathbb{X}^T\mathbb{X})^{-1}$$
$$= [(\mathbb{X}^T\mathbb{X})^{-1}(\mathbb{X}^T\mathbb{X})](\mathbb{X}^T\mathbb{X})^{-1}$$
$$= \mathbb{I}\,(\mathbb{X}^T\mathbb{X})^{-1}$$
$$= (\mathbb{X}^T\mathbb{X})^{-1}$$

$$\mathrm{Var}(\hat{\beta}) = \mathrm{Var}(\beta + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon)$$
$$= \mathrm{Var}((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon)$$
$$= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T \cdot \sigma^2\mathbb{I} \cdot \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}$$
$$= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}$$

using $\mathbb{A} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$

$$\mathbb{A}^T = \mathbb{X}\left[(\mathbb{X}^T\mathbb{X})^{-1}\right]^T$$
$$= \mathbb{X}\left[(\mathbb{X}^T\mathbb{X})^T\right]^{-1}$$
$$= \mathbb{X}\left[\mathbb{X}^T\mathbb{X}\right]^{-1}$$

$b = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T y$, $\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$

$\hat{\beta}$ is β plus some error due to the measurement process. →

$$\hat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(\mathbb{X}\beta + \varepsilon)$$
$$= \beta + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\varepsilon$$

Standard errors for coefficients in the linear model

- The formula $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}$ needs extra work to be useful for data analysis.
- In practice, we know the model matrix $\mathbb{X}$ but we don't know the measurement standard deviation σ.
- An estimate of the measurement standard deviation is the sample standard deviation of the residuals.
- For $\mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e}$ with $\mathbb{X}$ being $n \times p$, an estimate of σ is

$$s = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} = \sqrt{\frac{1}{n-p}\sum_{i=1}^{n}\left(y_i - [\mathbb{X}\mathbf{b}]_i\right)^2}$$

- We will discuss later why we choose to divide by $n - p$.
- The **standard error** of b_k for $k = 1, \ldots, p$ is

$$\mathrm{SE}(b_k) = s\sqrt{\left[\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\right]_{kk}}$$

which is an estimate of $\sqrt{\left[\mathrm{Var}(\hat{\boldsymbol{\beta}})\right]_{kk}}$.

- These standard errors are calculated by `lm()` in R.

k^{th} diagonal entry of $(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1}$

e.g. if $(\mathbb{X}^{\mathrm{T}}\mathbb{X})^{-1} = [V_{ij}]_{p\times p}$, the k^{th} diagonal entry is V_{kk}

```
lm1 <- lm(L_detrended~U_detrended) ; summary(lm1)

##
## Call:
## lm(formula = L_detrended ~ U_detrended)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.55654 -0.48641 -0.01867  0.40856  1.63118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.28999    0.09343   3.104  0.00281 **
## U_detrended  0.13137    0.06322   2.078  0.04161 *
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7705 on 66 degrees of freedom
## Multiple R-squared:  0.06141,Adjusted R-squared:  0.04718
## F-statistic: 4.318 on 1 and 66 DF,  p-value: 0.04161
```

How does R obtain linear model standard errors?

- The previous slide shows output from our analysis of unemployment and mortality from Chapter 1.
- Let's first extract the estimates and their standard errors from R, a good step toward reproducible data analysis.

```
names(summary(lm1))

##  [1] "call"          "terms"         "residuals"
##  [4] "coefficients"  "aliased"       "sigma"
##  [7] "df"            "r.squared"     "adj.r.squared"
## [10] "fstatistic"    "cov.unscaled"

summary(lm1)$coefficients

##                Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 0.2899928 0.09343146 3.103802 0.002812739
## U_detrended 0.1313673 0.06321939 2.077959 0.041606370
```

Extracting the design matrix

```
X <- model.matrix(lm1)
head(X)

##    (Intercept) U_detrended
## 16           1  -1.0075234
## 17           1   1.1027941
## 18           1   0.4881116
## 19           1  -1.5349043
## 20           1  -1.8662535
## 21           1  -2.0059360
```

Computing the standard errors directly

```
s <- sqrt(sum(resid(lm1)^2)/(nrow(X)-ncol(X))) ; s

## [1] 0.7704556

V <- s^2 * solve(t(X)%*%X)
sqrt(diag(V))

## (Intercept) U_detrended
##  0.09343146  0.06321939
```

- This matches the standard errors generated by `lm()`.

```
summary(lm1)$coefficients

##                Estimate Std. Error  t value    Pr(>|t|)
## (Intercept) 0.2899928 0.09343146 3.103802 0.002812739
## U_detrended 0.1313673 0.06321939 2.077959 0.041606370
```

Extracting the coefficient variance/covariance matrix

- The fitted `lm` object in R stores the estimated variance/covariance matrix for the coefficients in the output of `summary()`.

```
s <- summary(lm1)$sigma
XX <- summary(lm1)$cov.unscaled
s^2 * XX
```

```
##             (Intercept) U_detrended
## (Intercept) 0.008729439 0.000000000
## U_detrended 0.000000000 0.003996692
```

- This matches what we get from calculating $s^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}$ directly.

```
X <- model.matrix(lm1)
sum(resid(lm1)^2)/(nrow(X)-ncol(X)) * solve(t(X)%*%X)
```

```
##               (Intercept)  U_detrended
## (Intercept) 8.729439e-03 1.305064e-20
## U_detrended 1.305064e-20 3.996692e-03
```