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.

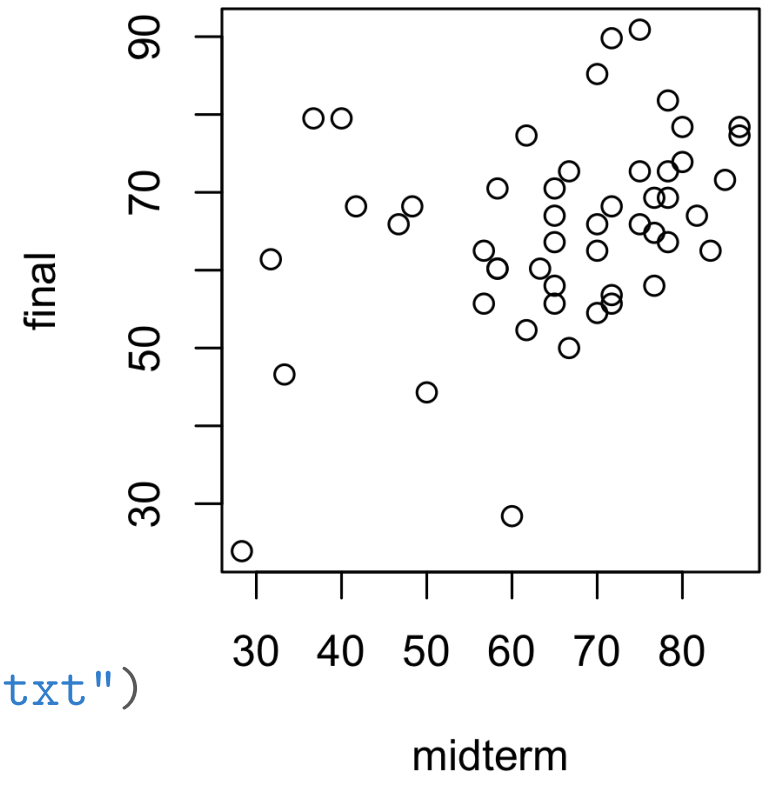
Example: midterm and final scores

- 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

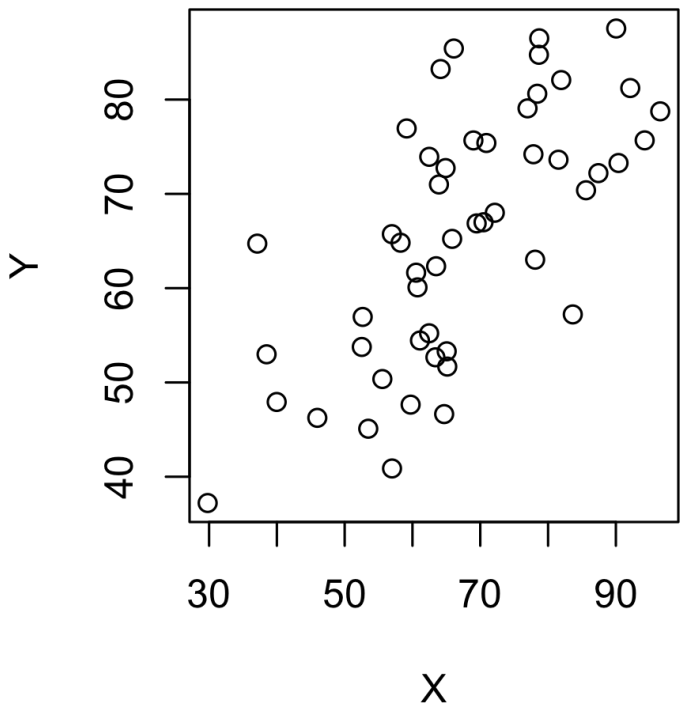
$$\text{Cov}(X, Y) = \text{E}\Big[\big(X - \text{E}[X]\,\big)\,\big(Y - \text{E}[Y]\,\big)\Big]$$

- 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**,

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

- 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



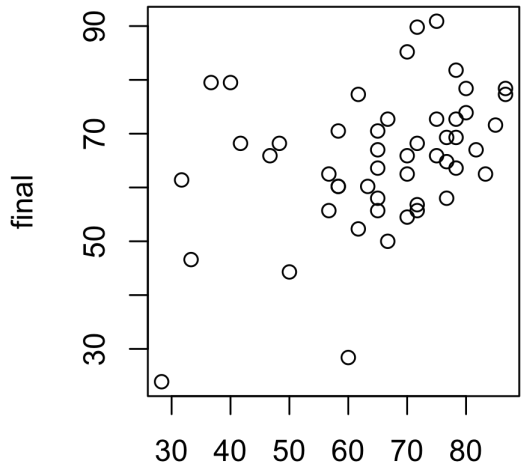

```
library(mvtnorm)
mvn <- rmvnorm(n=50,
  mean=c(X=65,Y=65),
  sigma=matrix(
    c(200,100,100,150),
    2,2)
)
plot(Y~X,data=mvn)
```

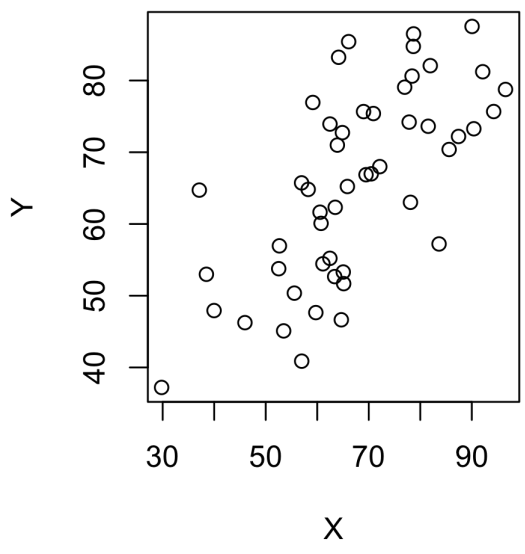
- We write $(X, Y) \sim \mathrm{MVN}(\boldsymbol{\mu}, \mathbb{V})$, where MVN is read "multivariate normal".

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

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?





More on covariance

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

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

- 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

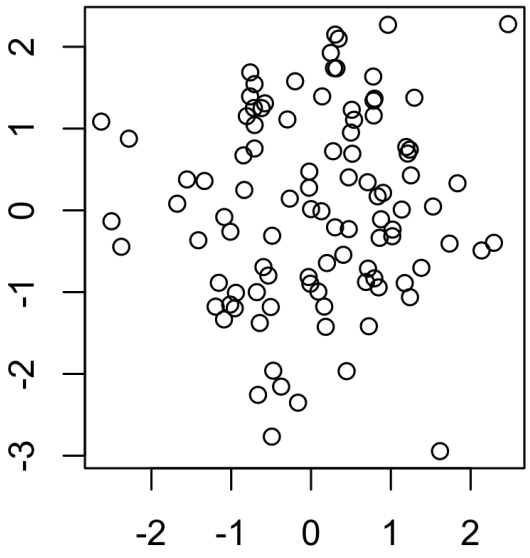
```
var(x$midterm)
## [1] 218.2155
var(x$final)
## [1] 169.7518
cov(x$midterm,x$final)
## [1] 75.61269
```

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

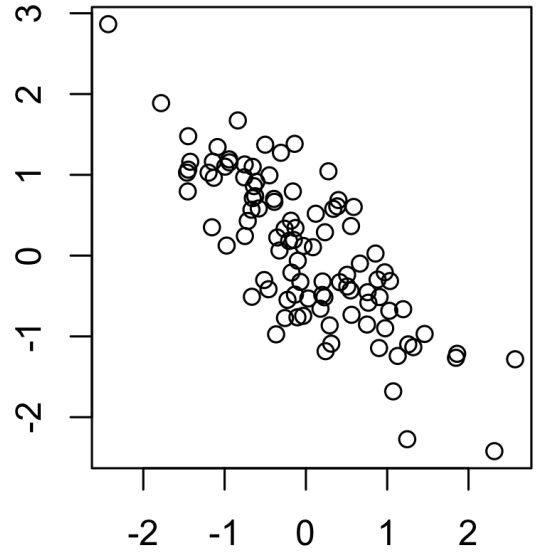
$$\mathrm{Cor}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} \qquad \mathrm{cor}(\mathbf{x},\mathbf{y}) = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{var}(\mathbf{x})\mathrm{var}(\mathbf{y})}}$$

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

```
rho <- 0
```

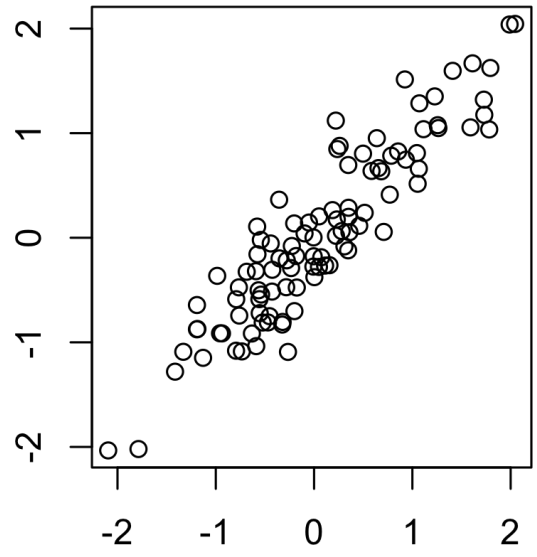



```
rho <- -0.8
```




```
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 **linearly dependent**.
- 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 $\mathrm{cov}(\mathbf{x}, \mathbf{y}) = 0.8$. What is the sample correlation, $\mathrm{cor}(\mathbf{x}, \mathbf{y})$?

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 $\mathrm{Cor}(X,Y)$.

More properties of covariance

- 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

(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()‘.

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.

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

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

```
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.

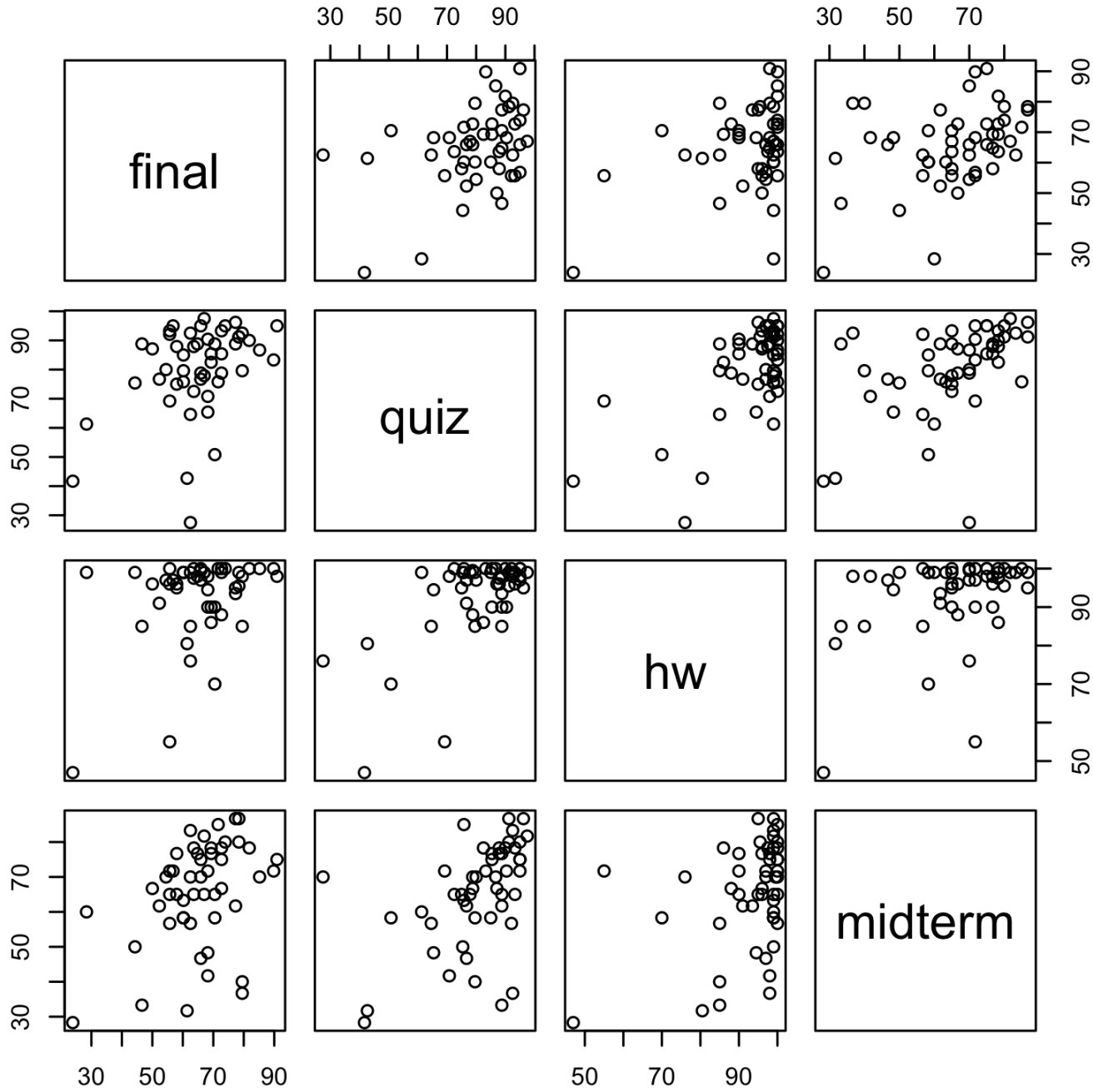
$$\text{cor}(\mathbb{X}) = \begin{bmatrix} 1 & \text{cor}(\mathbf{x}_1, \mathbf{x}_2) & \ldots & \text{cor}(\mathbf{x}_1, \mathbf{x}_p) \\ \text{cor}(\mathbf{x}_2, \mathbf{x}_1) & 1 & & \text{cor}(\mathbf{x}_2, \mathbf{x}_p) \\ \vdots & & \ddots & \vdots \\ \text{cor}(\mathbf{x}_p, \mathbf{x}_1) & \text{cor}(\mathbf{x}_p,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

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

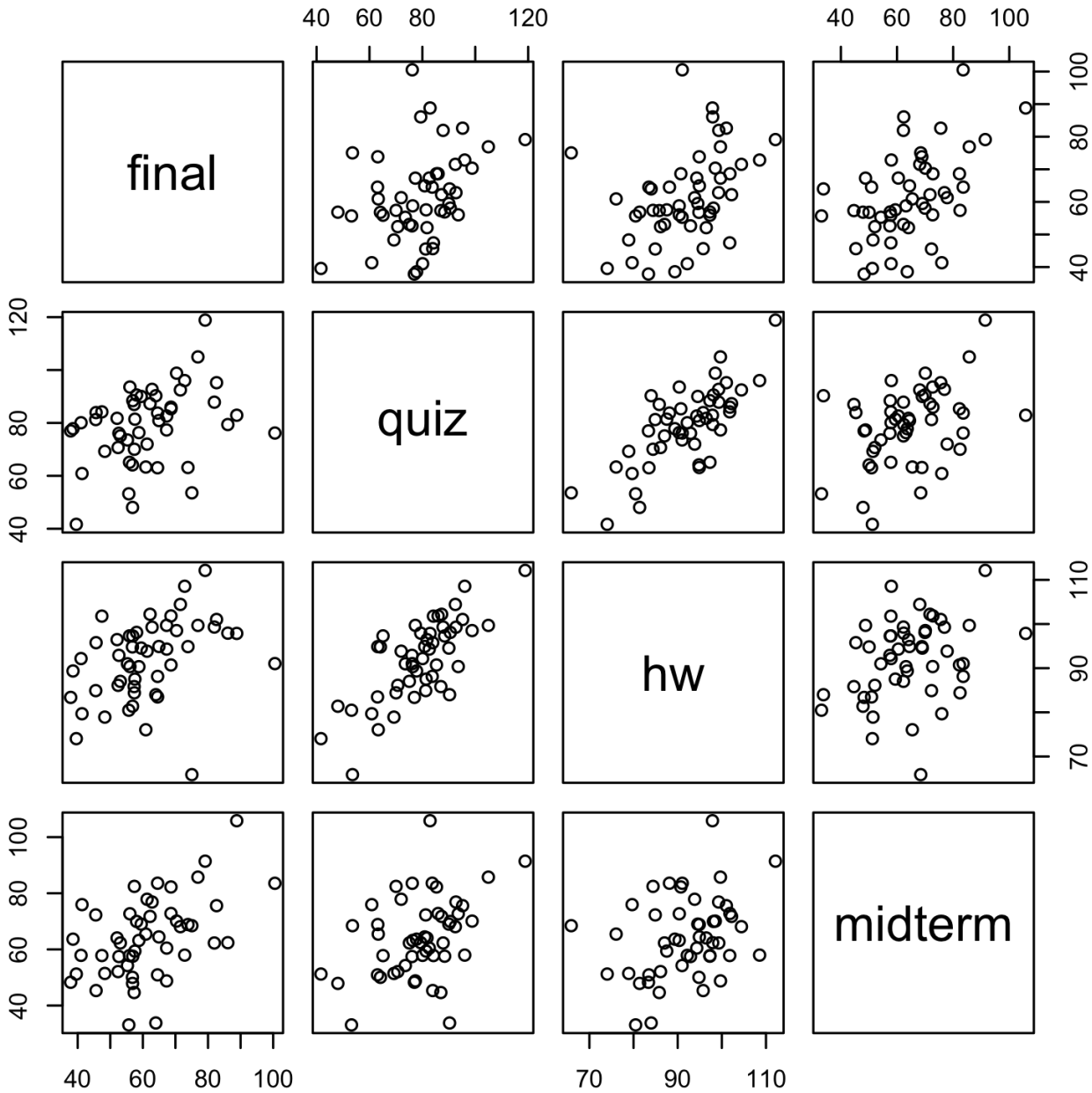
```
cor(x)
```

```
##              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)
```



```
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?

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?

Linear combinations

- 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.

Question 5.7. How would you construct a simulated data matrix $\mathbb{Z}_{\mathrm{sim}}$ from n realizations $\mathbf{Z}_1, \ldots, \mathbf{Z}_n$ of the random column vector $\mathbf{Z} = \mathbb{A}\mathbf{X}$? Be careful with transposes and keep track of dimensions.

Solution:

- There is a useful matrix variance/covariance formula for a linear combination.

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

Question 5.8. Add dimensions to each term in these equations to check they make sense.

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.

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

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

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

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

##           [,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.

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

- 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

- 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.

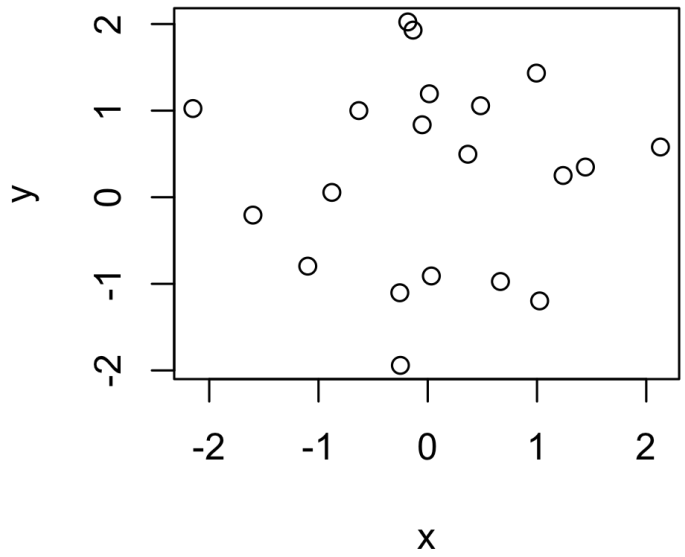
Zero correlation with and without independence

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

```r
cor(x,y)

## [1] 0.01825057
```

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

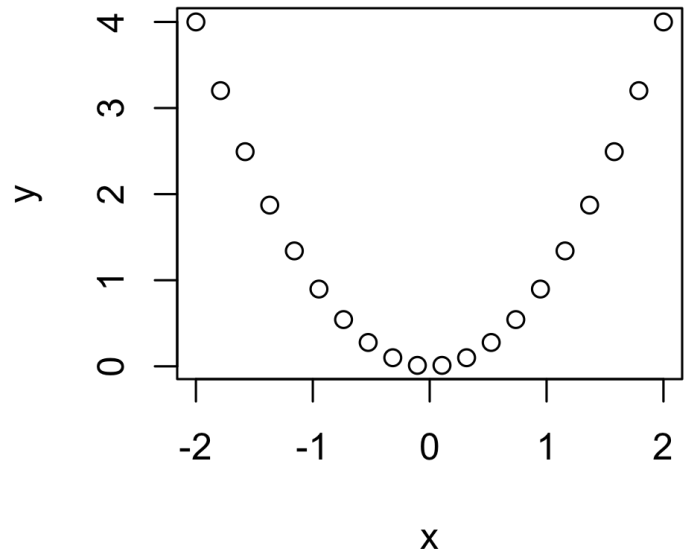



```r
x <- seq(-2,2,length=20)
y <- x^2
```

```r
cor(x,y)

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

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



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

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

where $\mathbf{0} = (0, \ldots, 0)$ and $\mathbb{I}$ is the $n \times n$ identity matrix. The off-diagonal entries of $\mathrm{Var}(\boldsymbol{\epsilon})$ are zero since $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$. For measurement error models, we break our usual rule of using upper case letters for random variables.

Example. A population version of the linear model

- First recall the sample version, which is

$$\text{(LM3)} \qquad \mathbf{y} = \mathbb{X}\,\mathbf{b} + \mathbf{e},$$

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 resulting vector of residuals.

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

$$\boxed{\text{(LM6)} \qquad \mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}}$$

where $\mathbb{X}$ is the same explanatory matrix as in (LM3), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is an unknown coefficient vector (we don't know the true population coefficient!) and $\boldsymbol{\epsilon}$ is measurement error with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$.

- Our model (LM6) asserts that the process which generated the response data $\mathbf{y}$ was like drawing a random vector $\mathbf{Y}$ consructed using a random measurement error model with known matrix $\mathbb{X}$ for some fixed but unknown value of $\boldsymbol{\beta}$.

Motivation for finding the means and variances of linear combinations of random variables

- Recall that the main purpose of having a probability model is so that we can investigate the chance variation due to picking the sample.
- Recall that for $(\mathrm{LM3})$, the least squares estimate is $\mathbf{b} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{y}$.
- This is a **statistic**, which means a function of the data and not a random variable. We cannot properly talk about the mean and variance of $\mathbf{b}$.
- We can work out the mean and variance of $\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$, as long as we know how to work out the mean and variance of linear combinations.
- As long as $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a **useful probability model** for the relationship between the response variable $\mathbf{y}$ and the explanatory variable $\mathbb{X}$, calculations done with this model may be useful.

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 be users and not abusers.

Example. For $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, we have $\boxed{\mathrm{E}[\mathbf{Y}] = \mathbb{X}\boldsymbol{\beta} + \mathrm{E}[\boldsymbol{\epsilon}] = \mathbb{X}\boldsymbol{\beta}}$

Example. For $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$, we have

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

- 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 centered around $\boldsymbol{\beta}$.

Linearity of expectation

- We have seen several versions of the same property that expectations can be moved through sums and multiplicative constants:

$$\begin{aligned}
\mathrm{E}[aX+b] &= a\mathrm{E}[X]+b, \\
\mathrm{E}\left[\sum_{i=1}^{n} a_i Y_i\right] &= \sum_{i=1}^{n} a_i \mathrm{E}[Y_i], \\
\mathrm{E}\left[\sum_{j=1}^{n} a_{ij} Y_j\right] &= \sum_{j=1}^{n} a_{ij} \mathrm{E}[Y_j]. \\
\mathrm{E}[\mathbb{A}\mathbf{Y}] &= \mathbb{A}\mathrm{E}[\mathbf{Y}]
\end{aligned}$$

- These properties are collectively known as **linearity**.
- Why? Maybe because these properties mean that linear equations for random variables lead to linear equations for their expectations.
- The linearity property means E follows a **distributive rule**. We can distribute E across sums just as we are used to doing in basic arithmetic.

Covariance of the least squares coefficients

- The covariance matrix formula we just developed can be written as

$$\mathrm{Var}(\mathbb{A}\mathbf{Y}) = \mathbb{A}\mathrm{Var}(\mathbf{Y})\mathbb{A}^{\mathrm{T}}.$$

Question 5.9. Consider the linear model $\mathbf{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\mathrm{E}[\boldsymbol{\epsilon}] = \mathbf{0}$ and $\mathrm{Var}(\boldsymbol{\epsilon}) = \sigma^2\mathbb{I}$. Apply this variance formula to $\hat{\boldsymbol{\beta}} = \left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}\mathbb{X}^{\mathrm{T}}\mathbf{Y}$ to get

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2\left(\mathbb{X}^{\mathrm{T}}\mathbb{X}\right)^{-1}$$

Standard errors for 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 error 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

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

- $SE(b_k)$ is an estimate of $\sqrt{\left[\mathrm{Var}(\hat{\boldsymbol{\beta}})\right]_{kk}}$.
- Let's check we now understand how `lm()` gets standard errors in R.

```r
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 SE 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

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
```