Foundations of Linear Regression

2. Review, Properties and Assumptions, Matrix Form

GOVT 6029 - Spring 2021

Cornell University

Outline

Announcements

- Office Hours

Announcements

- Office Hours
- Slide Deck

Announcements

- Office Hours
- Slide Deck
- Problem Set 1

Announcements

- Office Hours
- Slide Deck
- Problem Set 1
- Assigned next week, March 1. Due March 10

Announcements

- Office Hours
- Slide Deck
- Problem Set 1
- Assigned next week, March 1. Due March 10
- Questions, comments?

Outline

Outline

Guessing the correlation

Your turn

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

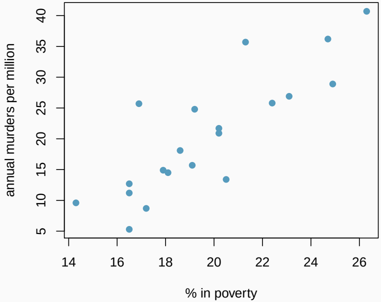
(a) -1.52

(b) -0.63

(c) -0.12

(d) 0.02

(e) 0.84



Your turn

Which of the following is the best guess for the correlation between annual murders per million and percentage living in poverty?

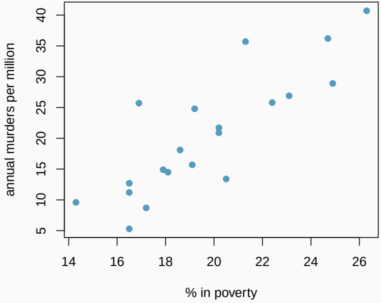
(a) -1.52

(b) -0.63

(c) -0.12

(d) 0.02

(e) *0.84*



Your turn

Which of the following is the best guess for the correlation between annual murders per million and population size?

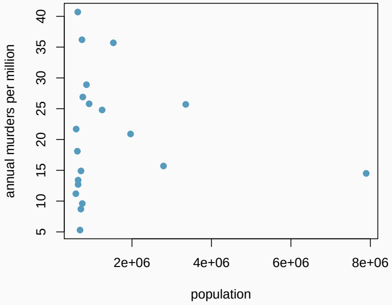
(a) -0.97

(b) -0.61

(c) -0.06

(d) 0.55

(e) 0.97



Your turn

Which of the following is the best guess for the correlation between annual murders per million and population size?

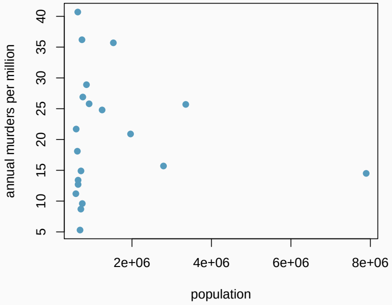
(a) -0.97

(b) -0.61

(c) *-0.06*

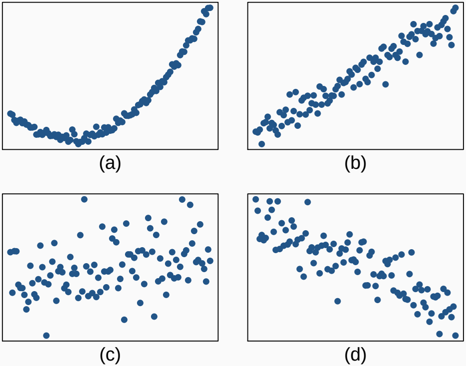
(d) 0.55

(e) 0.97



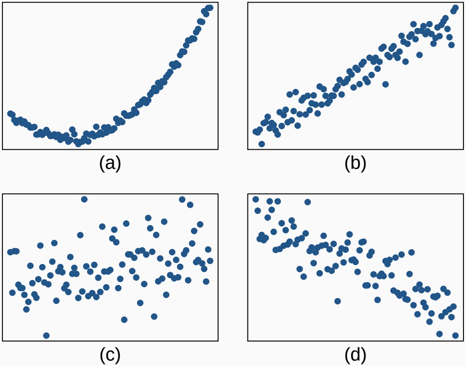
Your turn

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



Your turn

Which of the following is has the strongest correlation, i.e. correlation coefficient closest to +1 or -1?



(b) → correlation means linear association

Play the game!

http://guessthecorrelation.com/

Spurious correlations

Remember: correlation does not always imply causation!

http://www.tylervigen.com/

Outline

(2) Least squares line minimizes squared residuals

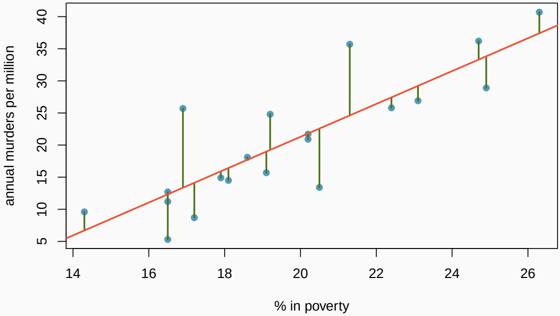
- Residuals are the leftovers from the model fit, and calculated as the difference between the observed and predicted y: $e_i = y_i - \hat{y}_i$

(2) Least squares line minimizes squared residuals

- Residuals are the leftovers from the model fit, and calculated as the difference between the observed and predicted y: $e_i = y_i - \hat{y}_i$
- The least squares line minimizes squared residuals:
    - Population data: $\hat{y} = \beta_0 + \beta_1 x$
    - Sample data: $\hat{y} = b_0 + b_1 x$

(2) Least squares line minimizes squared residuals

- Residuals are the leftovers from the model fit, and calculated as the difference between the observed and predicted y: $e_i = y_i - \hat{y}_i$
- The least squares line minimizes squared residuals:
    - Population data: $\hat{y} = \beta_0 + \beta_1 x$
    - Sample data: $\hat{y} = b_0 + b_1 x$



Outline

- *Slope:* For each unit increase in x, y is expected to behigher/lower on average by the slope.

$$b_1 = \frac{s_y}{s_x} R$$

Interpreting the last squares line

- *Slope:* For each unit increase in x, y is expected to behigher/lower on average by the slope.

$$b_1 = \frac{s_y}{s_x} R$$

- *Intercept:* When $x = 0$, y is expected to equal the intercept.

$$b_0 = \bar{y} - b_1 \bar{x}$$

- *Slope:* For each unit increase in x, y is expected to behighter/lower on average by the slope.

$$b_1 = \frac{s_y}{s_x} R$$

- *Intercept:* When $x = 0$, y is expected to equal the intercept.

$$b_0 = \bar{y} - b_1 \bar{x}$$

  - The calculation of the intercept uses the fact the a regression line **always** passes through $(\bar{x}, \bar{y})$.

Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

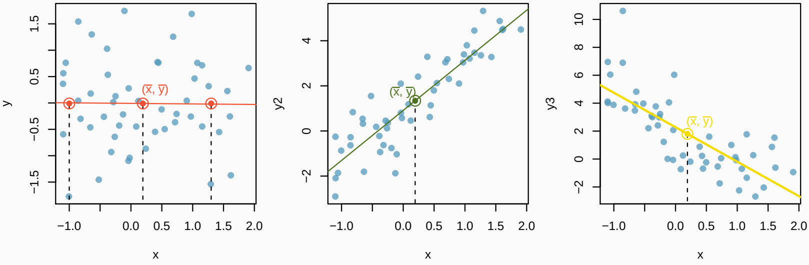
- If there is no relationship between x and y ($b_1 = 0$), the best guess for $\hat{y}$ for any value of x is $\bar{y}$.

Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

- If there is no relationship between x and y ($b_1 = 0$), the best guess for $\hat{y}$ for any value of x is $\bar{y}$.
- Even when there is a relationship between x and y ($b_1 \neq 0$), the best guess for $\hat{y}$ when $x = \bar{x}$ is still $\bar{y}$.

Why does the regression line **always** pass through $(\bar{x}, \bar{y})$?

- If there is no relationship between x and y ($b_1 = 0$), the best guess for $\hat{y}$ for any value of x is $\bar{y}$.
- Even when there is a relationship between x and y ($b_1 \neq 0$), the best guess for $\hat{y}$ when $x = \bar{x}$ is still $\bar{y}$.



Your turn

What is the interpretation of the slope?

$$\widehat{murders} = -29.91 + 2.56\ poverty$$

(a) Each additional percentage in those living in poverty increases number of annual murders per million by 2.56.

(b) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be higher by 2.56 on average.

(c) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be lower by 29.91 on average.

(d) For each percentage increase annual murders per million, the percentage of those living in poverty is expected to be higher by 2.56 on average.

Your turn

What is the interpretation of the slope?

$$\widehat{murders} = -29.91 + 2.56\ poverty$$

(a) Each additional percentage in those living in poverty increases number of annual murders per million by 2.56.

(b) *For each percentage increase in those living in poverty, the number of annual murders per million is expected to be higher by 2.56 on average.*

(c) For each percentage increase in those living in poverty, the number of annual murders per million is expected to be lower by 29.91 on average.

(d) For each percentage increase annual murders per million, the percentage of those living in poverty is expected to be higher by 2.56 on average.

Outline

Your turn

Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. For which of the following districts would you be most comfortable with your prediction?

A district where % in poverty =

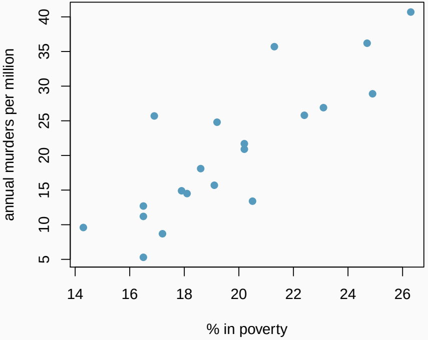
(a) 5%

(b) 15%

(c) 20%

(d) 26%

(e) 40%



Your turn

Suppose you want to predict annual murder count (per million) for a series of districts that were not included in the dataset. For which of the following districts would you be most comfortable with your prediction?

A district where % in poverty =

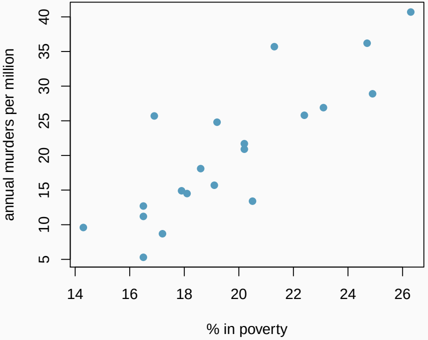
(a) 5%

(b) 15%

(c) *20%*

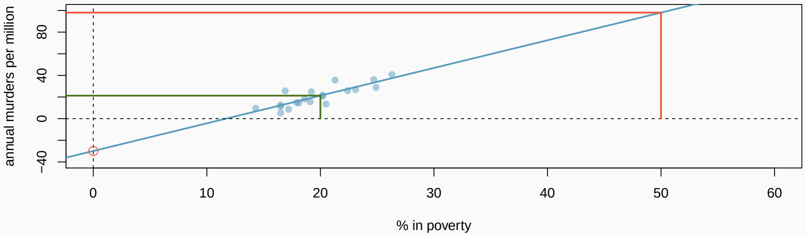
(d) 26%

(e) 40%



A note about the intercept

Sometimes the intercept might be an extrapolation: useful for adjusting the height of the line, but meaningless in the context of the data.



Calculating predicted values

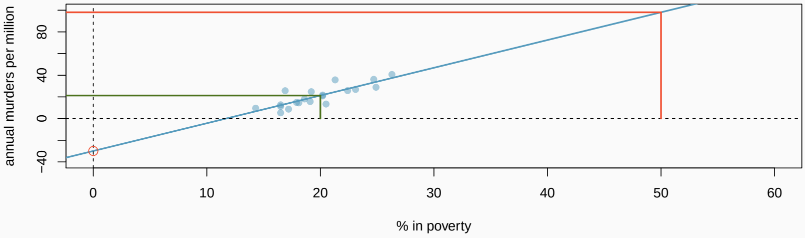
By hand: $\widehat{murder} = -29.91 + 2.56\ poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

By hand: $\widehat{murder} = -29.91 + 2.56\ poverty$

The predicted number of murders per million per year for a county with 20% poverty rate is:

$$\widehat{murder} = -29.91 + 2.56 \times 20 = 21.29$$



Outline

We can think of social science variables as comprised of two parts:

- **Systematic component**
  Determined by social relationships.
  e.g., part of your income is determined by your education.

We can think of social science variables as comprised of two parts:

- **Systematic component**
  Determined by social relationships.
  e.g., part of your income is determined by your education.
- **Stochastic component**
  Naturally occurring random variation.
  Not everyone with the same characteristics has the same income. Part of income is naturally random or stochastic, at least for practical purposes.

We can think of social science variables as comprised of two parts:

- **Systematic component**
  Determined by social relationships.
  e.g., part of your income is determined by your education.
- **Stochastic component**
  Naturally occurring random variation.
  Not everyone with the same characteristics has the same income. Part of income is naturally random or stochastic, at least for practical purposes.

Random variables contain both components

We can best understand random variables using probability distributions

Outline

A probability distribution describes in a random variable precisely in math

Suppose Y is a random variable. We can summarize it in two ways:

- pdf: probability density function, $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value

A probability distribution describes in a random variable precisely in math

Suppose Y is a random variable. We can summarize it in two ways:

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is less than or equal to that value

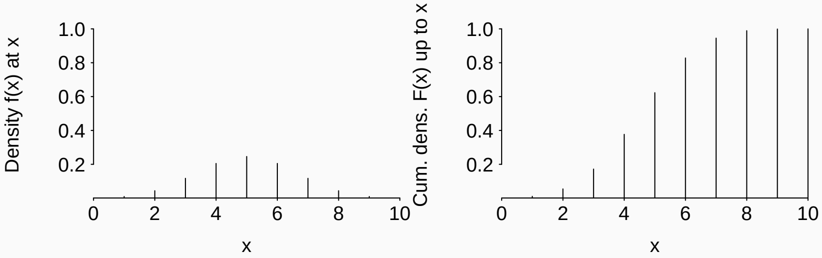
- pdf: probability density function, $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- cdf: cumulative density function, $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

If a variable can only take on certain values (eg, the number of children in a household), it has a discrete distribution:

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

If a variable can only take on certain values (eg, the number of children in a household), it has a discrete distribution:



- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

Note that the area under the pdf sums to 1,
while the cdf is the area under the pdf from the smallest possible value up to y

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

Note that the area under the pdf sums to 1,
while the cdf is the area under the pdf from the smallest possible value up to y

Thus, for discrete distributions, the cdf is the cumulative sum of the pdf:

$$F(Y) = \sum_{\forall Y \leq y} f(Y)$$

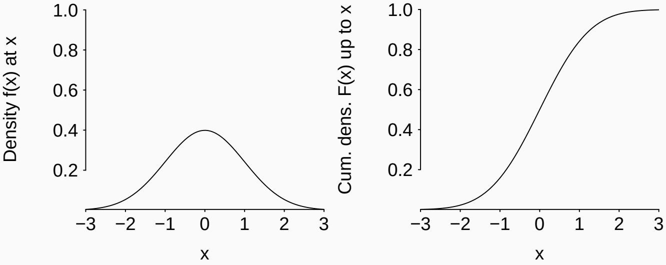
- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

If a variable can take on any (real) value, we must use a continuous distribution

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

If a variable can take on any (real) value, we must use a continuous distribution



- pdf: probability density function, $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is less than or equal to that value

Note that the area under the pdf sums to 1,
while the cdf is the area under the pdf from $-\infty$ to y

- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is less than or equal to that value

Note that the area under the pdf sums to 1,
while the cdf is the area under the pdf from $-\infty$ to y

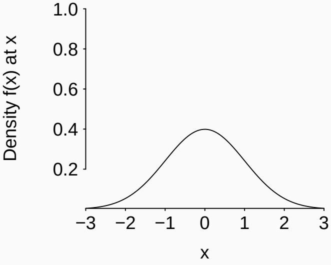
Thus for continuous distributions, the cdf is the integral of the pdf:

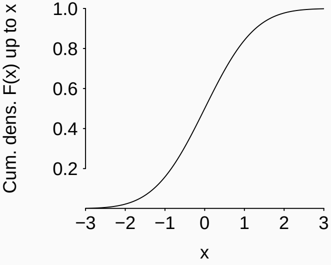
$$F(Y) = \int_{-\infty}^{y} f(Y)\mathrm{d}y$$

$$f_{\mathcal{N}}(y|\mu, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[\frac{-(y_i - \mu)^2}{2\sigma^2}\right]$$

Moments: $\mathrm{E}(y) = \mu \quad \mathrm{Var} = \sigma^2$

The Normal distribution is continuous and symmetric, with positive probability everywhere from $-\infty$ to ∞





What's the big deal about the Normal distribution?

One point of view: perhaps most continuous data are roughly Normally distributed

Why do people believe this?

They think the Central Limit Theorem applies to most data

Suppose we have N independent random variables $x_1, x_2, x_3, \ldots$.

Each x has an arbitrary probability distribution with mean μ_i and variance $\sigma_i^2 < \infty$

That is to say, these variables are not only independent, they could each have totally different distributions

Now suppose we average them all together into one super-variable,

$$X = \frac{1}{N} \sum_i x_i$$

The CLT shows that the distribution of this new variable, X, approaches a Normal distribution as $N \to \infty$

Proofs of the CLT are somewhat involved, so let's "verify" this by experiment

Flipping coins

The distribution of a coin flip is $\mathrm{Pr}(\mathrm{Heads}) = 0.5$,
$\mathrm{Pr}(\mathrm{Tails}) = 0.5$,
which is not bell-shaped at all

Suppose we flip M coins, and sum the number of heads.

If we repeat this exercise many times, the CLT says the resulting distribution of counts of heads should be approximately Normal.

For a proof and links on the CLT, see
`http://mathworld.wolfram.com/CentralLimitTheorem.htm`

Dropping balls

Dropping a ball through a pegboard mirrors the construction of a Normal random variable

Systematic component: the spot from which the balls are dropped

Stochastic component: the sum of all the random effects of the pegs

Result: a Normal distribution of ball locations

So why would many people think most continuous variables in the social sciences are Normal?

They are appealing to a "fuzzy" version of the CLT:

> *Data generated from many small and unrelated random shocks are approximately normally distributed*

One can see why, say, economic growth would be a good candidate for a Normally distributed variable

Application of the main tool introduced in this class, linear regression, usually based on this assumption

Review of simple linear regression

With the Normal distribution in mind, recall the linear regression model

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

ε_i is a normally distributed disturbance with mean 0 and variance σ^2

Equivalently, we write $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$

Note that:

The stochastic component has mean zero: $\mathrm{E}(\varepsilon_i) = 0$

The systematic component is: $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$

The errors are assumed uncorrelated: $\mathrm{E}(\varepsilon_i \times \varepsilon_j) = 0$ for all $i \neq j$

Review of simple linear regression

Recalling the definition of variance, note that in linear regression:

$$\begin{aligned}\sigma^2 &= \mathrm{E}\left((\varepsilon - \mathrm{E}(\varepsilon))^2\right)\\ &= \mathrm{E}\left((\varepsilon - 0)^2\right)\\ &= \mathrm{E}(\varepsilon^2)\end{aligned}$$

The square root of σ^2 is known as the standard error of the regression

It is how much we expect y to differ from its expected value, $\beta_0 + \beta_1 x_i$, on average

Linear Regression in Matrix Form

Scalar representation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

Equivalent matrix representation:

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times k}{\mathbf{X}} \; \underset{k \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\varepsilon}}$$

Writing out the matrices:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Linear Regression in Matrix Form

Note that we now have a vector of disturbances.

They have the same properties as before, but we will write them in matrix form.

The disturbances are still mean zero.

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \begin{bmatrix} \mathrm{E}(\varepsilon_1) \\ \mathrm{E}(\varepsilon_2) \\ \vdots \\ \mathrm{E}(\varepsilon_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Linear Regression in Matrix Form

But now we have an entire matrix of variances and covariances, $\mathbf{\Sigma}$

$$
\begin{aligned}
\mathbf{\Sigma} &= \begin{bmatrix} \mathrm{var}(\varepsilon_1) & \mathrm{cov}(\varepsilon_1, \varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_1, \varepsilon_n) \\ \mathrm{cov}(\varepsilon_2, \varepsilon_1) & \mathrm{var}(\varepsilon_2) & \dots & \mathrm{cov}(\varepsilon_2, \varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(\varepsilon_n, \varepsilon_1) & \mathrm{cov}(\varepsilon_n, \varepsilon_2) & \dots & \mathrm{var}(\varepsilon_n) \end{bmatrix} \\
&= \begin{bmatrix} \mathrm{E}(\varepsilon_1^2) & \mathrm{E}(\varepsilon_1\varepsilon_2) & \dots & \mathrm{E}(\varepsilon_1\varepsilon_n) \\ \mathrm{E}(\varepsilon_2\varepsilon_1) & \mathrm{E}(\varepsilon_2^2) & \dots & \mathrm{E}(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\varepsilon_n\varepsilon_1) & \mathrm{E}(\varepsilon_n\varepsilon_2) & \dots & \mathrm{E}(\varepsilon_n^2) \end{bmatrix}
\end{aligned}
$$

However, the above matrix can be written far more compactly as an outer product

$$\mathbf{\Sigma} = \boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'$$

Linear Regression in Matrix Form

Recall $\mathrm{E}(\varepsilon_i \varepsilon_j) = 0$ for all $i \neq j$,
so all of the off-diagonal elements above are zero by assumption

Recall also that all ε_i are assumed to have the same variance, σ^2

So *if* the linear regression assumptions hold,
the variance-covariance matrix has a simple form:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 \mathbf{I}$$

When these assumptions do not hold,
we will need more complex models than simple linear regression

Linear Regression in Matrix Form

So how do we solve for $\boldsymbol{\beta}$?

Let's use the least squares principle:
choose $\hat{\boldsymbol{\beta}}$ such that the sum of the squared errors is minimized

In symbols, we want

$$\arg\min_{\boldsymbol{\beta}} \sum_i \varepsilon_i^2 \qquad \text{or, in matrix form} \qquad \arg\min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$

This is a straightforward minimization (calculus) problem.
The trick is using matrices to simplify notation.

The sum of squared errors can be written out as

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

(what is this notation doing? why do we need the transpose?)

Linear Regression in Matrix Form

We need two bits of matrix algebra:

$$
\begin{aligned}
(\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}' \\
\left(\begin{bmatrix} 10 \\ 3 \end{bmatrix}+\begin{bmatrix} 2 \\ 6 \end{bmatrix}\right)' &= \begin{bmatrix} 10 & 3 \end{bmatrix}+\begin{bmatrix} 2 & 6 \end{bmatrix} \\
\begin{bmatrix} 12 & 9 \end{bmatrix} &= \begin{bmatrix} 12 & 9 \end{bmatrix}
\end{aligned}
$$

and

$$
\begin{aligned}
(\mathbf{X}\boldsymbol{\beta})' &= \boldsymbol{\beta}'\mathbf{X}' \\
\begin{bmatrix} 2 & 1 \\ 5 & 6 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} &= \begin{bmatrix} 3 & 4 \end{bmatrix}\begin{bmatrix} 2 & 5 \\ 1 & 6 \end{bmatrix} \\
\begin{bmatrix} (2\times 3)+(1\times 4) \\ (5\times 3)+(6\times 4) \end{bmatrix}' &= \begin{bmatrix} (3\times 2)+(4\times 1) & (3\times 5)+(4\times 6) \end{bmatrix} \\
\begin{bmatrix} 10 & 39 \end{bmatrix} &= \begin{bmatrix} 10 & 39 \end{bmatrix}
\end{aligned}
$$

Linear Regression in Matrix Form

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

First, we distribute the transpose:

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y}' - (\mathbf{X}\boldsymbol{\beta})')(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Next, let's substitute $\boldsymbol{\beta}'\mathbf{X}'$ for $(\mathbf{X}\boldsymbol{\beta})'$

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y}' - \boldsymbol{\beta}'\mathbf{X})(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Multiplying this out, we get

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \mathbf{y}'\mathbf{y} - \beta'\mathbf{X}'\mathbf{y} - \mathbf{y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Simplifying, we get

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Linear Regression in Matrix Form

Now we need to take the derivative with respect to $\boldsymbol{\beta}$,
to see which $\boldsymbol{\beta}$ minimize the sum of squares.

How do we take the derivative of a scalar with respect to a vector?

It's just a bunch of scalar derivatives stacked together:

$$\frac{\partial y}{\partial \mathbf{x}} = \left[\begin{array}{cccc} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \ldots & \frac{\partial y}{\partial x_n} \end{array} \right]'$$

For example, for $\mathbf{a}$ and $\mathbf{x}$ both $n \times 1$ vectors

$$\begin{aligned} y = \mathbf{a}'\mathbf{x} &= a_1 x_1 + a_2 x_2 + \ldots + a_n x_n \\ \frac{\partial y}{\partial \mathbf{x}} &= \left[\begin{array}{cccc} a_1 & a_2 & \ldots & a_n \end{array} \right]' \\ \frac{\partial y}{\partial \mathbf{x}} &= \mathbf{a} \end{aligned}$$

Linear Regression in Matrix Form

A similar pattern holds for quadratic expresssions.

Note the vector analogue of x^2 is the inner product $\mathbf{x}'\mathbf{x}$

And the vector analogue of ax^2 is $\mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times n$ matrix of coefficients

$$
\begin{aligned}
\frac{\partial a x^2}{\partial x} &= 2ax \\
\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} &= 2\mathbf{A}\mathbf{x}
\end{aligned}
$$

The details are a bit more complicated ($\mathbf{x}'\mathbf{A}\mathbf{x}$ is the sum of a lot of terms), but the intuition is the same.

Linear Regression in Matrix Form

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

Taking the derivative of this expression, and setting it equal to 0, we get

$$\frac{\partial \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = 0$$

This is a mimimum,
and the $\boldsymbol{\beta}$'s that solve this equation thus minimize the sum of squares.

So let's solve for $\boldsymbol{\beta}$:

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

This is the least squares estimator for $\boldsymbol{\beta}$

As long as we have software to help us with matrix inversion, it is easy to calculate.