Foundations of Linear Regression

2. Review, Properties and Assumptions, Matrix Form

GOVT 6029 - Spring 2019

Cornell University

Outline

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

Announcements

- Problem Set 1 ... Questions, comments?

Announcements

- Problem Set 1 ... Questions, comments?
- Due Friday html and Rmd

Announcements

- Problem Set 1 ... Questions, comments?
- Due Friday html and Rmd
- Canvas

Outline

1. Housekeeping

2. Main ideas

   1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables
   2. Least squares line minimizes squared residuals
   3. Interpreting the least squares line
   4. Predict, but don't extrapolate
   5. Regression in social sciences
   6. Distributions
   7. Linear Regression in Matrix Form

3. Assumptions

Outline

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables
    2. Least squares line minimizes squared residuals
    3. Interpreting the least squares line
    4. Predict, but don't extrapolate
    5. Regression in social sciences
    6. Distributions
    7. Linear Regression in Matrix Form

3. Assumptions

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?

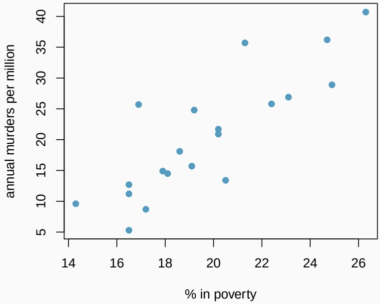
(a) -1.52

(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?

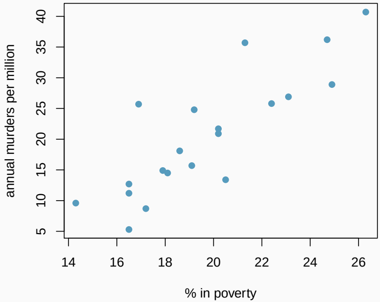
(a) -1.52

(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?

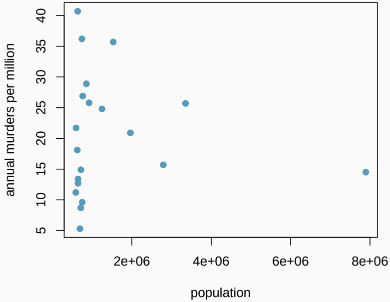
(a) -0.97

(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?

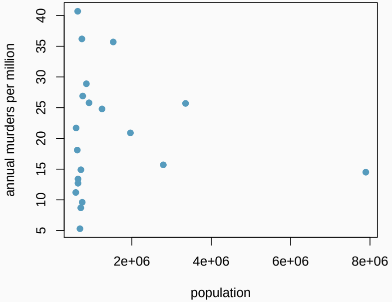
(a) -0.97

(b) -0.61

(c) **-0.06**

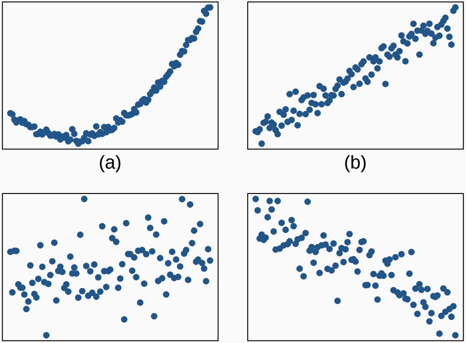
(d) 0.55

(e) 0.97



Your turn

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

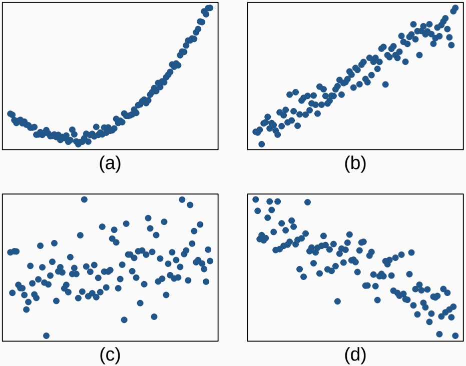


(a) (b) (c) (d)

Assessing the correlation

Your turn

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



(b) $\rightarrow$ correlation means <u>linear</u> association

Play the game!

http://guessthecorrelation.com/

Spurious correlations

Remember: correlation does not always imply causation!

http://www.tylervigen.com/

Outline

1. Housekeeping

2. Main ideas

   1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

   2. Least squares line minimizes squared residuals

   3. Interpreting the least squares line

   4. Predict, but don't extrapolate

   5. Regression in social sciences

   6. Distributions

   7. Linear Regression in Matrix Form

3. Assumptions

(2) Least squares line minimizes squared residuals

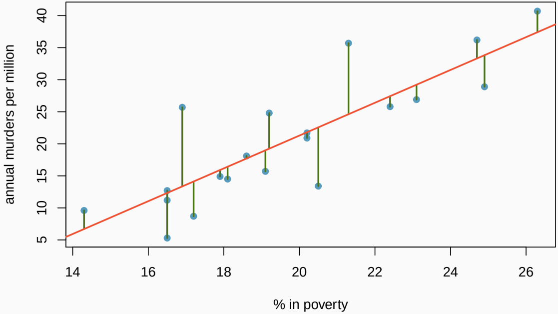
- Residuals are the leftovers from the model fit, and calculated as the difference between the observed and predicted y:
  $e_i = y_i - \hat{y}_i$

(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

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

Interpreting the last squares line

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

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

- *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 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.

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

  - 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 $(\bar{x}, \bar{y})$?

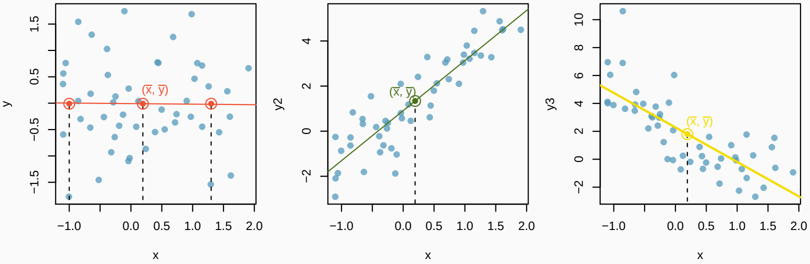
- If there is no relationship between x and y ($b_1 = 0$), the best guess for $\hat{y}$ for any value of x is $\bar{y}$.

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

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

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 =

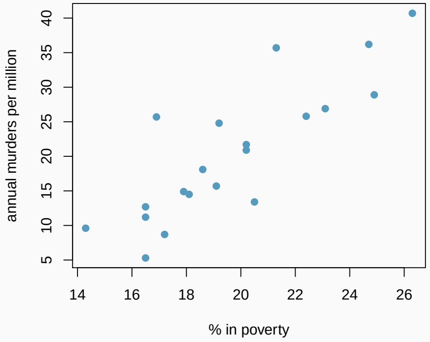
(a) 5%

(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 =

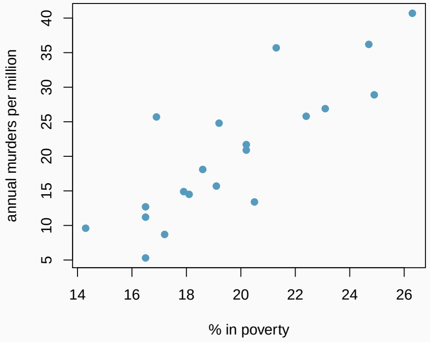
(a) 5%

(b) 15%

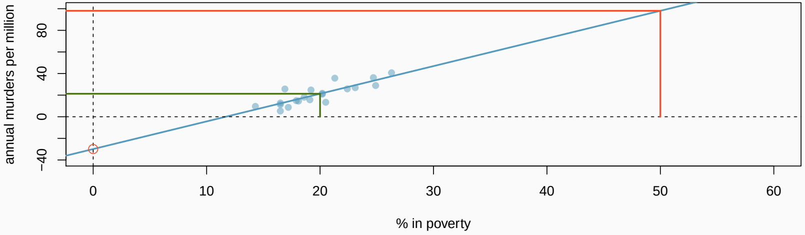
(c) **20%**

(d) 26%

(e) 40%



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

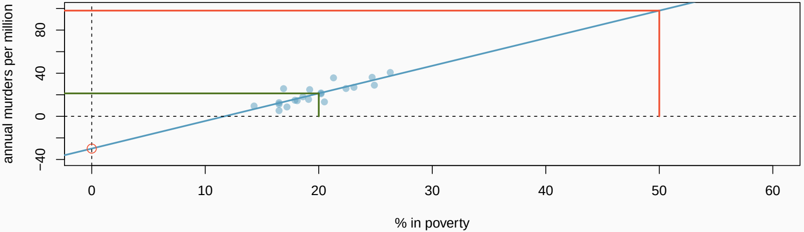
By hand: $\widehat{murder} = -29.91 + 2.56\ poverty$

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

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

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

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

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

Probability distributions

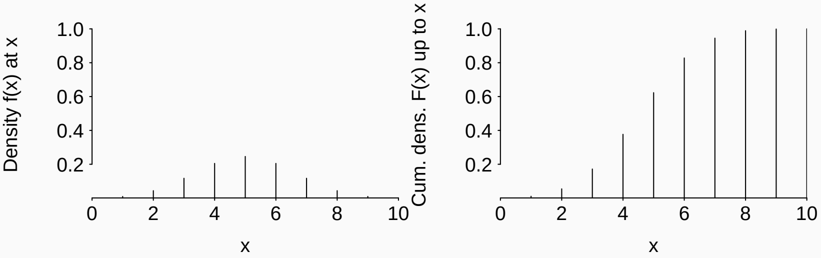
- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

- **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:

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

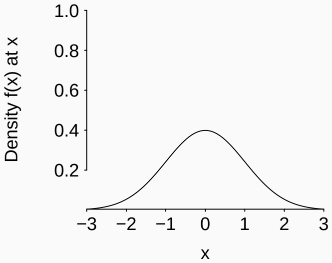
- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

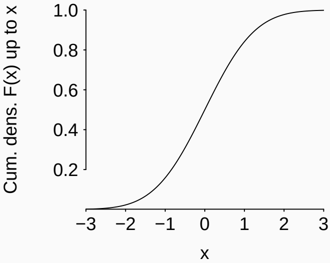
- **pdf: probability density function,** $f(Y)$
  For any possible value of $Y = y$, gives the probability that Y takes on that value
- **cdf: cumulative density function,** $F(Y)$
  For any possible value of $Y = y$,
  gives the probability that Y is $<=$ to that value

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





Probability distributions

- **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:

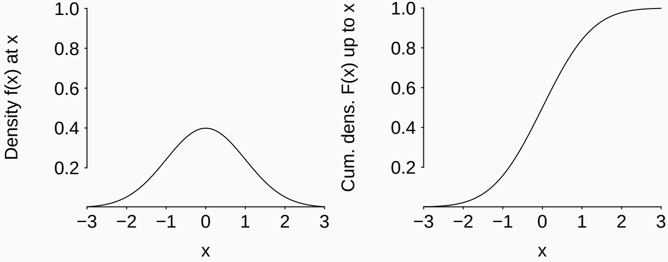
$$F(Y) = \int_{-\infty}^{y} f(Y) \mathrm{d}y$$

The Normal (Gaussian) distribution

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

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

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



The Normal distribution

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 $N \to \infty$

Outline

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

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: $\mathrm{E}(\varepsilon_i) = 0$

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: $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$

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 $i \neq j$

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

$$\sigma^2 = \mathrm{E}\left((\varepsilon - \mathrm{E}(\varepsilon))^2\right)$$

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

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

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:

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

Linear Regression in Matrix Form

Scalar representation:

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

Equivalent matrix representation:

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

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:

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

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.

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

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

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

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

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

$$= \begin{bmatrix} \mathrm{E}(\varepsilon_1^2) & \mathrm{E}(\varepsilon_1\varepsilon_2) & \dots & \mathrm{E}(\varepsilon_1\varepsilon_n) \\ \mathrm{E}(\varepsilon_2\varepsilon_1) & \mathrm{E}(\varepsilon_2^2) & \dots & \mathrm{E}(\varepsilon_2\varepsilon_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}(\varepsilon_n\varepsilon_1) & \mathrm{E}(\varepsilon_n\varepsilon_2) & \dots & \mathrm{E}(\varepsilon_n^2) \end{bmatrix}$$

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

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

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 <u>if</u> the linear regression assumptions hold,
the variance-covariance matrix has a simple form:

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

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

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

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

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

$$\arg\min_{\boldsymbol{\beta}} \sum_i \varepsilon_i^2$$

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.

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

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

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

Linear Regression in Matrix Form

We need two bits of matrix algebra:

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$$

We need two bits of matrix algebra:

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

We need two bits of matrix algebra:

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

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

$$(\mathbf{X}\beta)' = \beta'\mathbf{X}'$$

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

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

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

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

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

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

Linear Regression in Matrix Form

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

Linear Regression in Matrix Form

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

First, we distribute the transpose:

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

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

First, we distribute the transpose:

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

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

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

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

First, we distribute the transpose:

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

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

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

Multiplying this out, we get

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

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

First, we distribute the transpose:

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

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

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

Multiplying this out, we get

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

Simplifying, we get

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

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:

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

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:

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \frac{\partial y}{\partial x_2} & \dots & \frac{\partial y}{\partial x_n} \end{bmatrix}'$$

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

$$\begin{aligned} y = \mathbf{a}'\mathbf{x} &= a_1x_1 + a_2x_2 + \ldots + a_nx_n \\ \frac{\partial y}{\partial \mathbf{x}} &= \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}' \\ \frac{\partial y}{\partial \mathbf{x}} &= \mathbf{a} \end{aligned}$$

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

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

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

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.

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

Take the derivative of the expression, setting it $= 0$, we get

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

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

Linear Regression in Matrix Form

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

Take the derivative of the expression, setting it $= 0$, we get

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

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

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

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

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

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.

Outline

1. Housekeeping

2. Main ideas

    1. Correlation coefficient describes the strength and direction of the linear association between two numerical variables

    2. Least squares line minimizes squared residuals

    3. Interpreting the least squares line

    4. Predict, but don't extrapolate

    5. Regression in social sciences

    6. Distributions

    7. Linear Regression in Matrix Form

3. Assumptions

What makes an estimator good?

Is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ a good estimate of β?

Would another estimator be better?

What would an alternative be?

Maybe minimizing the sum of absolute errors?

Or something nonlinear?

First we'll have to decide what makes an estimator good.

What makes an estimator good?

Some common criteria:

What makes an estimator good?

Some common criteria:

Bias

Some common criteria:

Bias The meaning of bias in statistics is more specific (and at times at variance) with plain English.

It does not mean subjectivity.

What makes an estimator good?

Some common criteria:

Bias The meaning of bias in statistics is more specific (and at times at variance) with plain English.

It does not mean subjectivity.

Is the estimate $\hat{\beta}$ provided by the model expected to equal the true value β?

If not, how far off is it?

What makes an estimator good?

Some common criteria:

Bias The meaning of bias in statistics is more specific (and at times at variance) with plain English.

It does not mean subjectivity.

Is the estimate $\hat{\boldsymbol{\beta}}$ provided by the model expected to equal the true value $\boldsymbol{\beta}$?

If not, how far off is it?

This is the **bias**, $\mathrm{E}(\hat{\beta} - \beta)$

Some common criteria:

Bias The meaning of bias in statistics is more specific (and at times at variance) with plain English.

It does not mean subjectivity.

Is the estimate $\hat{\boldsymbol{\beta}}$ provided by the model expected to equal the true value $\boldsymbol{\beta}$?

If not, how far off is it?

This is the **bias**, $\mathrm{E}(\hat{\beta} - \beta)$

Although it seems "obvious" on face that we always prefer an unbiased estimator if one is available we also want the estimate to be close to the truth most of the time

What makes an estimator good?

Unbiased methods are not perfect.

They usually still miss the truth by some amount,
But the direction in which they miss is not systematic or known ahead of time.

Unbiased estimates can be useless.

What makes an estimator good?

Unbiased methods are not perfect.

They usually still miss the truth by some amount,
But the direction in which they miss is not systematic or known ahead of time.

Unbiased estimates can be useless. Why?

One unbiased estimate of the time of day:

What makes an estimator good?

Unbiased methods are not perfect.

They usually still miss the truth by some amount,
But the direction in which they miss is not systematic or known ahead of time.

Unbiased estimates can be useless. Why?

One unbiased estimate of the time of day:a random draw from the numbers 0–24. Utterly useless.

Unbiased methods are not perfect.

They usually still miss the truth by some amount,
But the direction in which they miss is not systematic or known ahead of time.

Unbiased estimates can be useless. Why?

One unbiased estimate of the time of day:a random draw from the numbers 0–24. Utterly useless.

Biased estimates are not necessarily terrible.

A biased estimate of the time of day: a clock that is 2 minutes fast.

What makes an estimator good?

Efficiency:

Efficiency: Efficient estimators get closest to the truth on average

What makes an estimator good?

Efficiency: Efficient estimators get closest to the truth on average

Measures of efficiency answer the question:
How much do we miss the truth by on average?

Efficiency thus incorporates both bias and variance of estimator.

What makes an estimator good?

Efficiency: Efficient estimators get closest to the truth on average

Measures of efficiency answer the question:
How much do we miss the truth by on average?

Efficiency thus incorporates both bias and variance of estimator.

A biased est with low variance may be “better” than an unbiased high var est

Efficiency: Efficient estimators get closest to the truth on average

Measures of efficiency answer the question:
How much do we miss the truth by on average?

Efficiency thus incorporates both bias and variance of estimator.

A biased est with low variance may be “better” than an unbiased high var est

Efficiency: Efficient estimators get closest to the truth on average

Some examples:

Efficiency: Efficient estimators get closest to the truth on average

Some examples:

	Unbiased?	Efficient?
Stopped clock.		
Random clock.		
Clock that is “a lot fast”		
Clock that is “a little fast”		
A well-run atomic clock		

Efficiency: Efficient estimators get closest to the truth on average

Some examples:

	Unbiased?	Efficient?
Stopped clock.	No	No
Random clock.	Yes	No
Clock that is “a lot fast”	No	No
Clock that is “a little fast”	No	Yes
A well-run atomic clock	Yes	Yes

What makes an estimator good?

To measure efficiency, we use **mean squared error**:

$$\begin{aligned} \text{MSE} &= \mathrm{E}\left[\left(\beta - \hat{\beta}\right)^2\right] \\ &= \mathrm{Var}(\hat{\beta}) + \mathrm{Bias}(\hat{\beta}|\beta)^2 \end{aligned}$$

$\sqrt{MSE}$ is how much you miss the truth by on average

In most cases, we want to use the estimator that minimizes MSE
We will be especially happy when this is also an unbiased estimator
But it won't always be

What makes an estimator good?

Consistency:

What makes an estimator good?

Consistency: An estimator that converges to the truth as the number of observations grows

What makes an estimator good?

Consistency: An estimator that converges to the truth as the number of observations grows

Formally, $\mathrm{E}(\hat{\beta} - \beta) \to 0$ as $N \to \infty$

Of great concern to many econometricians

Not as great a concern in political science

What makes an estimator good?

Consistency: An estimator that converges to the truth as the number of observations grows

Formally, $\mathrm{E}(\hat{\beta} - \beta) \to 0$ as $N \to \infty$

Of great concern to many econometricians

Not as great a concern in political science (as a thought experiment, $N \to \infty$ doesn't help much when the observations are, say, industrialized countries)

What makes an estimator good?

Consistency: An estimator that converges to the truth as the number of observations grows

Formally, $\mathrm{E}(\hat{\beta} - \beta) \to 0$ as $N \to \infty$

Of great concern to many econometricians

Not as great a concern in political science (as a thought experiment, $N \to \infty$ doesn't help much when the observations are, say, industrialized countries)

We will be mainly concerned with efficiency, secondarily with bias, and hardly at all with consistency

What can go wrong in the linear model?

Two things that can go wrong:

- omitted variable bias
- specification bias

What can go wrong in the linear model?

Two things that can go wrong:

- omitted variable bias
- specification bias

Another important source of bias is **selection**:

If we select observations non-randomly from the world, we may get biased estimates of means, regression coefficients, and other quantities

Another important source of bias is **selection**:

If we select observations non-randomly from the world, we may get biased estimates of means, regression coefficients, and other quantities

Another important source of bias is **selection**:

If we select observations non-randomly from the world, we may get biased estimates of means, regression coefficients, and other quantities

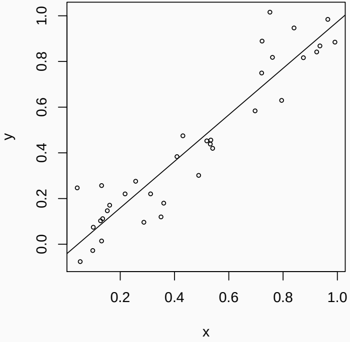
- Average children per marriage is 2.5. How many were in your family growing up? Are these numbers different? Who is "left out" in the second sample?

Another important source of bias is **selection**:

If we select observations non-randomly from the world, we may get biased estimates of means, regression coefficients, and other quantities

- Average children per marriage is 2.5. How many were in your family growing up? Are these numbers different? Who is "left out" in the second sample?
- In testimony to NY state senate, motorcyclists testified that in their (multiple) crashes, helmets would not have prevented injuries. Who didn't testify?
- Regression example: Selection on the observed variables

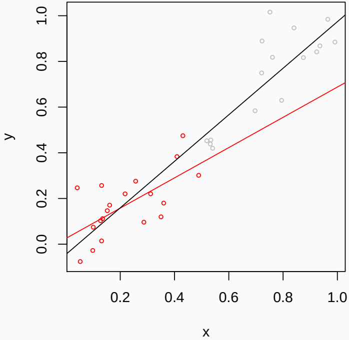
Selection bias



Suppose we conducted a survey & asked people their income (x) and conservatism (y)

With the full range of respondents, we find a strong relationship

Selection bias

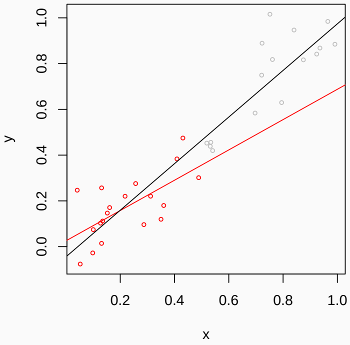


But suppose high income (or highly conservative) people decline to answer

Then we run a regression on the red dots only.

And get a result biased towards 0.

Selection bias



$\rightarrow$ Try to maximize variance of covariates, and avoid selecting on response variables

Most selection is unintentional, so think hard about sources of selection bias

What else can go wrong in a linear regression?

Even if your data are sampled without bias from the population of interest, and your model correctly specified, several data problems can violate the linear regression assumptions

What else can go wrong in a linear regression?

Even if your data are sampled without bias from the population of interest, and your model correctly specified, several data problems can violate the linear regression assumptions

In order of declining severity:

Perfect collinearity

Endogeneity of covariates

Heteroskedasticity

Serial correlation

Non-normality

Lots of new jargon. Let's work through it.

Perfect Collinearity

Perfect collinearity occurs when $\mathbf{X}'\mathbf{X}$ is singular; ie, the determinant $|\mathbf{X}'\mathbf{X}| = 0$

Happens when two or more columns of $\mathbf{X}$ are linearly dependent on each other

Perfect Collinearity

Perfect collinearity occurs when $\mathbf{X}'\mathbf{X}$ is singular; ie, the determinant $|\mathbf{X}'\mathbf{X}| = 0$

Happens when two or more columns of $\mathbf{X}$ are linearly dependent on each other

Common causes: including a variable twice, or a variable and itself times a constant

Perfect Collinearity

Perfect collinearity occurs when $\mathbf{X}'\mathbf{X}$ is singular; ie, the determinant $|\mathbf{X}'\mathbf{X}| = 0$

Happens when two or more columns of $\mathbf{X}$ are linearly dependent on each other

Common causes: including a variable twice, or a variable and itself times a constant

Very rare—except in panel data, as we will see

Matrix inversion—and thus LS regression—is impossible here

“Partial” Collinearity

What if our covariates are correlated but not perfectly so?

"Partial" Collinearity

What if our covariates are correlated but not perfectly so?

Then they are not linearly dependent

"Partial" Collinearity

What if our covariates are correlated but not perfectly so?

Then they are not linearly dependent

The regression coefficients are identified (a unique estimate exists for each β)

“Partial” Collinearity

What if our covariates are correlated but not perfectly so?

Then they are <u>not</u> linearly dependent

The regression coefficients are identified (a unique estimate exists for each β)

Regression with partial collinearity is unbiased & efficient.

But if the correlation among the X's is high, there is little to distinguish them

This leads to noisy estimates and large standard errors

“Partial” Collinearity

What if our covariates are correlated but not perfectly so?

Then they are not linearly dependent

The regression coefficients are identified (a unique estimate exists for each β)

Regression with partial collinearity is unbiased & efficient.

But if the correlation among the X’s is high, there is little to distinguish them

This leads to noisy estimates and large standard errors

Those large se’s are *correct*

"Partial" Collinearity is actually an oxymoron.

Inappropriately, this situation is sometimes called "multicollinearity"

“Partial” Collinearity is actually an oxymoron.

Inappropriately, this situation is sometimes called “multicollinearity”

Technically, multicollinearity describes only perfect linear dependence

Linear regression does not “fail” when correlation among **X** is “high”.

“Partial” Collinearity

“Partial” Collinearity is actually an oxymoron.

Inappropriately, this situation is sometimes called “multicollinearity”

Technically, multicollinearity describes only perfect linear dependence

Linear regression does not “fail” when correlation among **X** is “high”.

There is no “fix” for high correlation: it is not a statistical problem.

Have highly correlated **X** and large se’s?

"Partial" Collinearity

"Partial" Collinearity is actually an oxymoron.

Inappropriately, this situation is sometimes called "multicollinearity"

Technically, multicollinearity describes only perfect linear dependence

Linear regression does not "fail" when correlation among **X** is "high".

There is no "fix" for high correlation: it is not a statistical problem.

Have highly correlated **X** and large se's?
Then you lack sufficient data to precisely answer your research question

So far, we have (implicitly) taken our regressors, $\mathbf{X}$, as fixed

$\mathbf{X}$ is not dependent on $\mathbf{Y}$

Fixed = pre-determined = exogenous

Exogenous & endogenous variables

So far, we have (implicitly) taken our regressors, $\mathbf{X}$, as fixed

$\mathbf{X}$ is not dependent on $\mathbf{Y}$

Fixed = pre-determined = exogenous

$\mathbf{Y}$ consists of a function of $\mathbf{X}$ plus an error

$\mathbf{Y}$ is thus endogenous to $\mathbf{X}$

endogenous = “determined within the system”

What if **Y** helps determine **X** in the first place?

That is, what if there is reciprocal causation?

Exogenous & endogenous variables

What if **Y** helps determine **X** in the first place?

That is, what if there is reciprocal causation?

Very common in political science:

- campaign spending and share of the popular vote.
- policy attitudes and party identification
- arms races and war, etc.
- exchange rate policy and inflation

Exogenous & endogenous variables

What if **Y** helps determine **X** in the first place?

That is, what if there is reciprocal causation?

Very common in political science:

- campaign spending and share of the popular vote.
- policy attitudes and party identification
- arms races and war, etc.
- exchange rate policy and inflation

In these cases, **Y** and **X** are both endogenous

Least squares is biased in this case

It will remain biased even as you add more data

In other words, it is <u>inconsistent</u>, or biased even as $N \to \infty$

Linear regression allows us to model the mean of a variable well

Y could be any linear function of $\boldsymbol{\beta}$ and $\mathbf{X}$

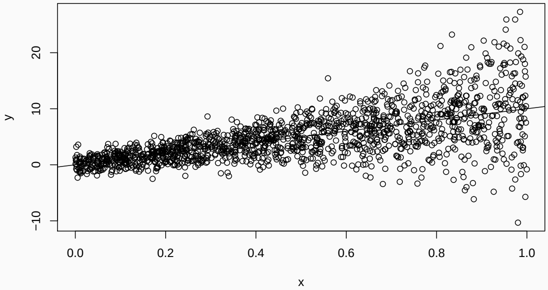
But LS always assumes the variance of that variable is the same:

σ^2, a constant

We don't think $\mathbf{Y}$ has constant mean. Why expect constant variance?

In fact, heteroskedasticity—non-constant error variance—is very common

Heteroskedasticity: “Different variance”

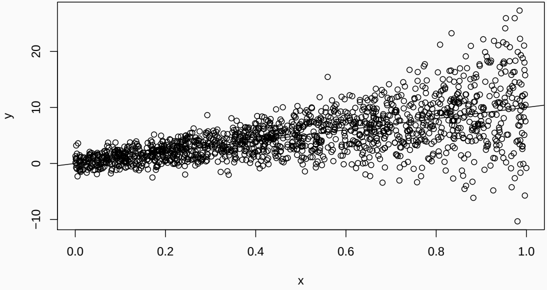


A common pattern of heteroskedasticity:

Variance and mean increase together

Here, they are both correlated with the covariate X

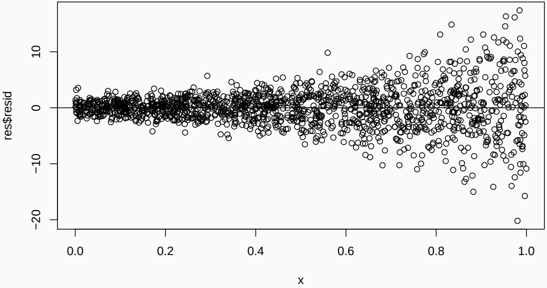
Heteroskedasticity: "Different variance"



In a fuzzy sense, X is a necessary but not sufficient condition for Y

This is usually an important point substantively. Heteroskedasticity is <u>interesting</u>, not just a nuisance

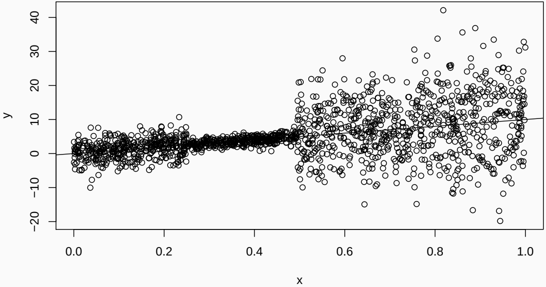
Heteroskedasticity: "Different variance"



We can usually find heteroskedasticity by plotting the residuals against each covariate

Look for a pattern. Often a megaphone

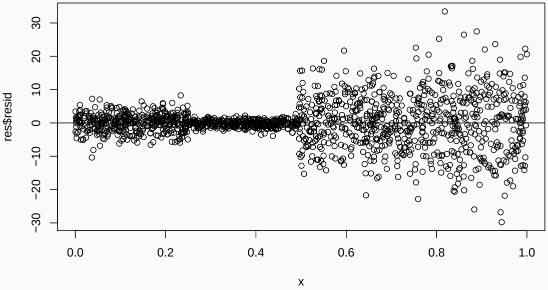
Heteroskedasticity: "Different variance"



But other patterns are possible.

Above, there is a dramatic difference in variance in different parts of the dataset

Heteroskedasticity: "Different variance"



The same diagnostic reveals this problem.

Heteroskedasticity of this type often appears in panel datasets, where there are groups of observations from different units that each share a variance

Unpacking σ^2

Every observation consists of a systematic component ($\mathbf{x}_i\boldsymbol{\beta}$) and a stochastic component (ε_i)

Generally, we can think of the stochastic component as an *n*-vector $\boldsymbol{\varepsilon}$ following a multivariate normal distribution:

$$\boldsymbol{\varepsilon} \sim \mathcal{MVN}(\mathbf{0}, \boldsymbol{\Sigma})$$

Aside: how the Multivariate Normal distribution works

Consider the simplest multivariate normal distribution, the joint distribution of two normal variables $\mathbf{x}_1$ and $\mathbf{x}_2$

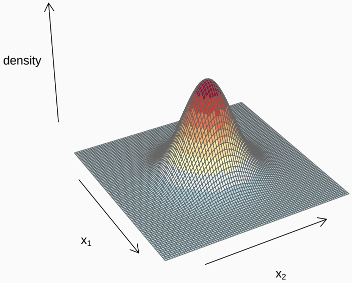
As usual, let μ indicate a mean, and σ a variance or covariance

$$\mathbf{X} = \mathcal{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix} \right)$$

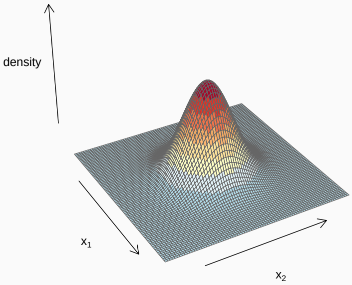
The MVN is more than the sum of its parts:
There is a mean and variance for each variable, <u>and</u> covariance between each pair

The Multivariate Normal distribution



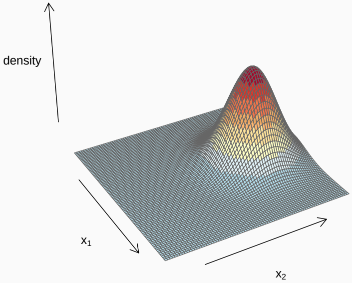
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



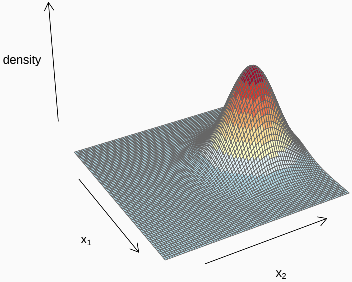
The standard MVN, with zero means, unit variances, and no covariance, looks like a higher dimension version of the normal: a symmetric mountain of probability

The Multivariate Normal distribution



$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

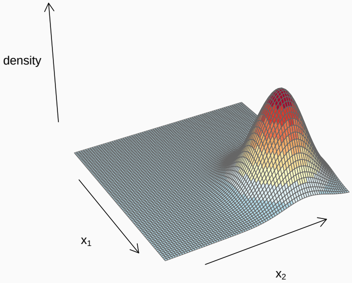
The Multivariate Normal distribution



Shifting the mean of $\mathbf{x}_2$ moves the MVN in one dimension only

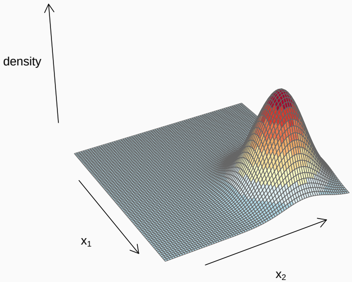
Mean shifts affect only one dimension at a time

The Multivariate Normal distribution



$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

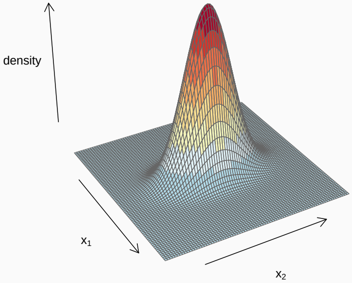
The Multivariate Normal distribution



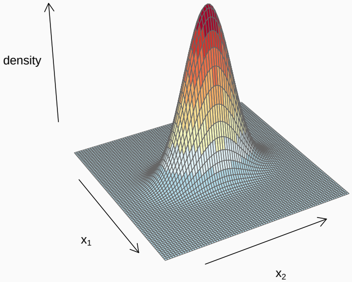
We could, of course, move the means of our variables at the same time.

This MVN says the most likely outcome is both x_1 and x_2 will be near 2.0

The Multivariate Normal distribution

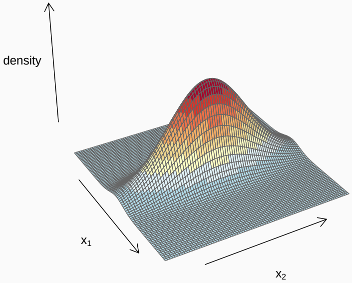


$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 1 \end{bmatrix} \right)$$



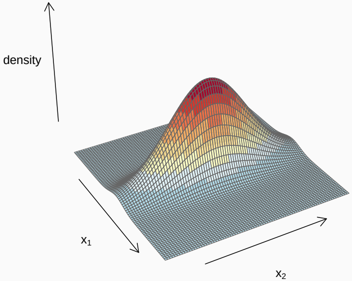
Shrinking the variance of $\mathbf{x}_1$ moves the mass of probability towards the mean of $\mathbf{x}_1$, but leaves the distribution around x_2 untouched

The Multivariate Normal distribution



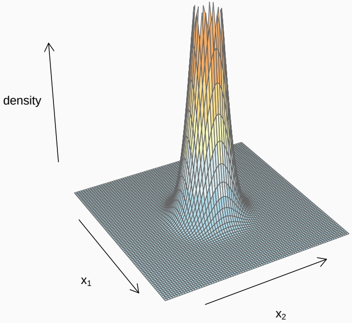
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 3 \end{bmatrix} \right)$$

The Multivariate Normal distribution

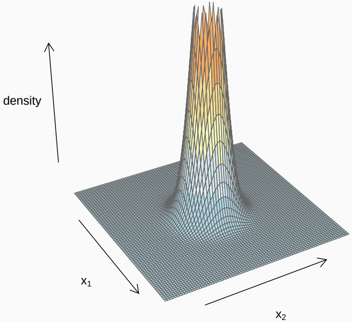


Increasing the variance of $\mathbf{x_2}$ spreads the probability out,
so we are less certain of $\mathbf{x_2}$, but just as certain of $\mathbf{x_1}$ as before

The Multivariate Normal distribution

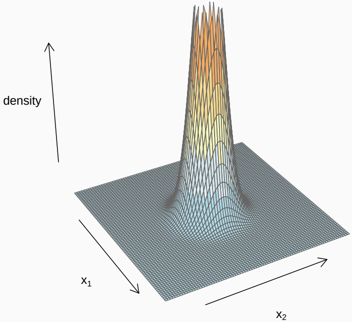


$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 0.33 \end{bmatrix} \right)$$



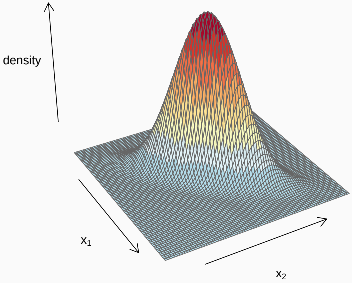
If the variance is small on all dimensions, the distribution collapses to a spike over the means of all variables

The Multivariate Normal distribution

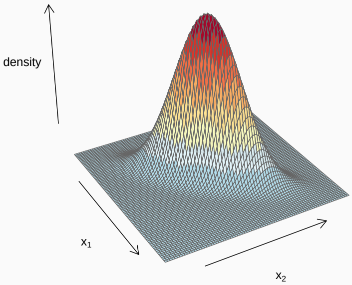


In this case, we are fairly certain of where all our variables tend to lie

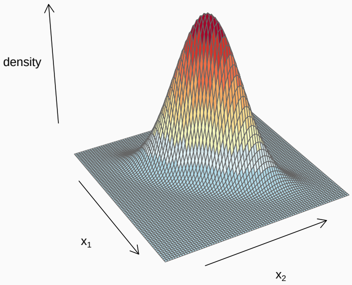
The Multivariate Normal distribution



$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$$

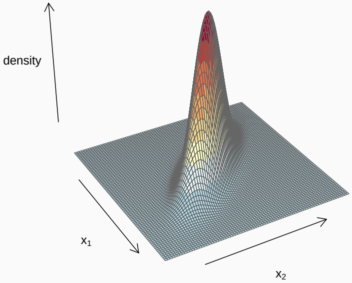


In this special case, with unit variances, the covariance is also the correlation, so our distribution say x_1 and x_2 are correlated at $r = 0.8$



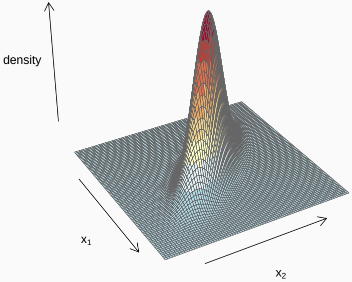
A positive correlation between our variables makes the MVN asymmetric,
with greater mass on likely combinations

The Multivariate Normal distribution



$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



A negative correlation makes mismatched values of our covariates more likely

In our current example, we have a huge multivariate normal distribtion:

each observation has its own mean and variance, and a covariance with every other observation

Suppose we have four observations. The Var-cov matrix of the disturbances is then

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

In its most "ordinary" form, linear regression puts strict conditions on the variance-variance matrix, $\boldsymbol{\Sigma}$

Again, assuming we have only four observations, the Var-cov matrix is

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

Could treat each observation as consisting of $\mathbf{x}_i\boldsymbol{\beta}$ and a separate, univariate normal disturbance, each with the same variance, σ^2.

This is the usual linear regression set up

Will look like our first example MVN: a symmetric mountain, but

Unpacking σ^2: heteroskedastic case

Suppose the distrurbances are heteroskedastic.

Now each observation has an error term drawn from a Normal with its own variance

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

Unpacking σ^2: heteroskedastic case

Suppose the distrurbances are heteroskedastic.

Now each observation has an error term drawn from a Normal with its own variance

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & 0 \\ 0 & 0 & 0 & \sigma_4^2 \end{bmatrix}$$

Still no covariance across disturbances.

Even so, we now have more parameters than we can estimate.

If every observation has its own unknown variance, we cannot estimate them

This MVN looks like the first example of a ridge: steeper in some directions than others, but not "tilted"

Unpacking σ^2: heteroskedastic case

Heteroskedasticity does not bias least squares

But LS is inefficient in the presence of heteroskedasticity

More efficient estimators give greater weight to observations with low variance

They pay more attention to the signal, and less attention to the noise

Unpacking σ^2: heteroskedastic case

Heteroskedasticity does not bias least squares

But LS is inefficient in the presence of heteroskedasticity

More efficient estimators give greater weight to observations with low variance

They pay more attention to the signal, and less attention to the noise

Heteroskedasticity tends to make se's incorrect, because they depend on the estimate of σ^2

Researchers often try to "fix" standard errors to deal with this

(more on this later)

Unpacking σ^2: heteroskedasticity & autocorrelation

Suppose each disturbance has its own variance, and may be correlated with other disturbances

The most general case allows for both heteroskedasticity & autocorrelation

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{bmatrix}$$

LS is unbiased but inefficient in this case

The standard errors will be wrong, however

Key application: time series.

Current period is usually a function of the past

Gauss-Markov Conditions

So when is least squares unbiased?

When is it efficient?

When are the standard errors correct?

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\varepsilon) = 0$	

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\varepsilon) = 0$	Biased, even as $N \to \infty$

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	**X** is exogenous	$\mathrm{E}(\mathbf{X}\varepsilon) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\varepsilon) = 0$	

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\varepsilon) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\varepsilon) = 0$	Biased, even as $N \to \infty$

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
4	No serial correlation	$\mathrm{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
4	No serial correlation	$\mathrm{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	Unbiased but ineff. se's wrong
5	Homoskedastic errors	$\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$	

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \rightarrow \infty$
3	Disturbances have mean 0	$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \rightarrow \infty$
4	No serial correlation	$\mathrm{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	Unbiased but ineff. se's wrong
5	Homoskedastic errors	$\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$	Unbiased but ineff. se's wrong
6	Gaussian error distrib	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$	

To judge the performance of LS, we'll need to make some assumptions

#	Assumption	Formal statement	Consequence of violation
1	No (perfect) collinearity	$\mathrm{rank}(\mathbf{X}) = k, k < n$	Coefficients unidentified
2	$\mathbf{X}$ is exogenous	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
4	No serial correlation	$\mathrm{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	Unbiased but ineff. se's wrong
5	Homoskedastic errors	$\mathrm{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$	Unbiased but ineff. se's wrong
6	Gaussian error distrib	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$	se's wrong unless $N \to \infty$

(Assumptions get stronger from top to bottom, but 4 & 5 could be combined)

It is easy to show β_{LS} is linear and unbiased, under Asps 1–3:

If $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$, then by substitution

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{LS}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \end{aligned}$$

So long as

- $(\mathbf{X}'\mathbf{X})^{-1}$ is uniquely identified,
- $\mathbf{X}$ is exogenous or at least uncorrelated with ε, and
- $E(\varepsilon) = 0$ (regardless of the distribution of ε)

Then $\mathbf{E}(\hat{\boldsymbol{\beta}}_{\mathrm{LS}}) = \boldsymbol{\beta}$

$\rightarrow$ β_{LS} is unbiased and a linear function of $\mathbf{y}$.

Gauss-Markov Theorem

If we make assumptions 1–5, we can make a stronger claim

When there is no serial correlation, no heteroskedasticity, no endogeneity, and no perfect collinearity, then

Gauss-Markov holds that LS is the best linear unbiased estimator (BLUE)

BLUE means that among linear estimators that are unbiased, $\hat{\beta}_{\mathrm{LS}}$ has the least variance.

But, there might be a nonlinear estimator with lower MSE overall, unless . . .

If in addition to Asp 1–5, the disturbances are normally distributed (6), then

Gauss-Markov holds LS is Minimum Variance Unbiased (MVU)