POLS/CSSS 503:
Advanced Quantitative Political Methodology

Linear Regression in Matrix Form / Properties & Assumptions of Linear Regression

Christopher Adolph

Department of Political Science
and
Center for Statistics and the Social Sciences
University of Washington, Seattle

Agenda

Quick review of linear regression model in scalar form

Linear regression in matrix form

Assumptions & properties of the linear regression model

Consequences of violations of those assumptions

Programming and simulation in `R`

Random Variables

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

Probability distributions

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

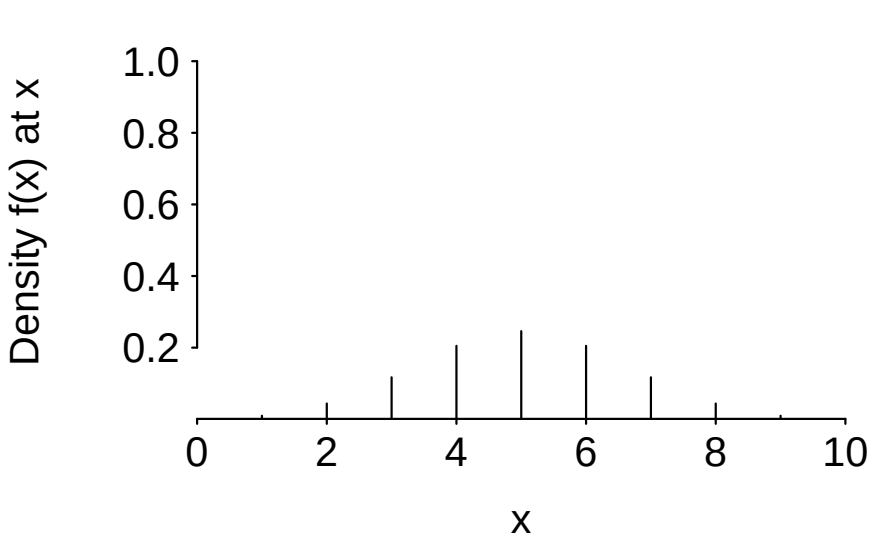
- **pdf: probability density function,** $f(Y)$

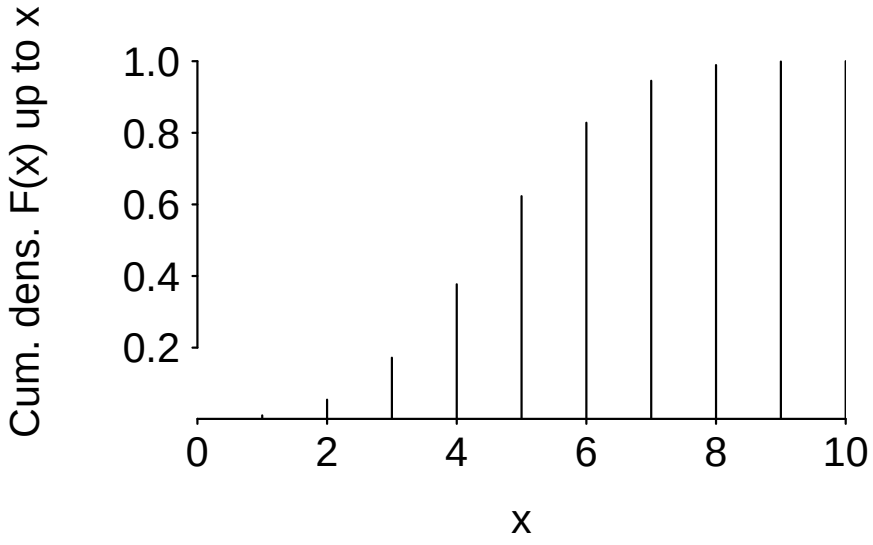
  For any possible value of $Y = y$, gives the probability that Y takes on that value

- **cdf: cumulative density function,** $F(Y)$

  For any possible value of $Y = y$,
  gives the probability that Y is less than or equal 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*:





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 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 the smallest possible value up to y

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

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

Probability distributions

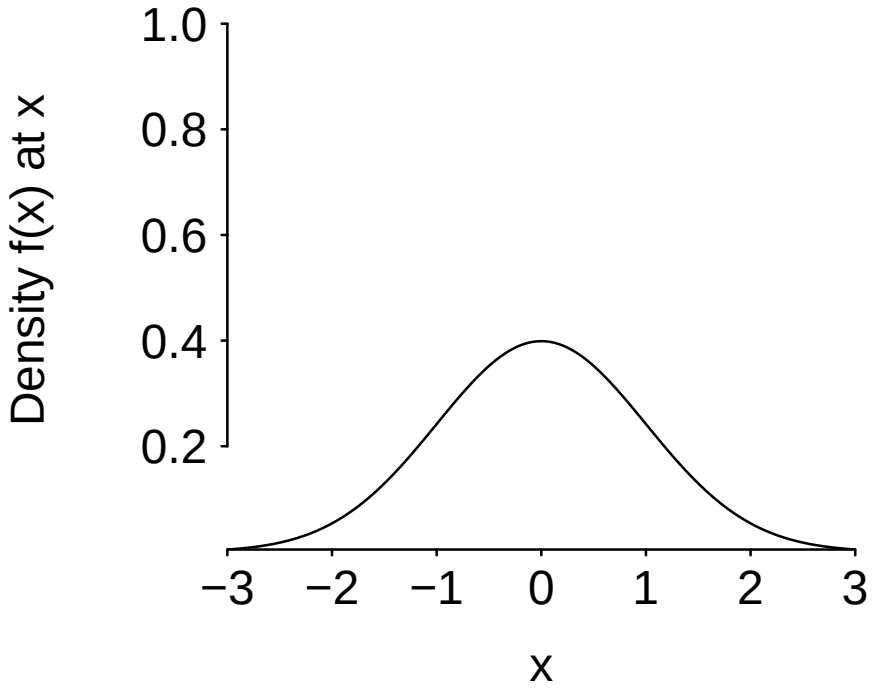
- **pdf: probability density function,** $f(Y)$

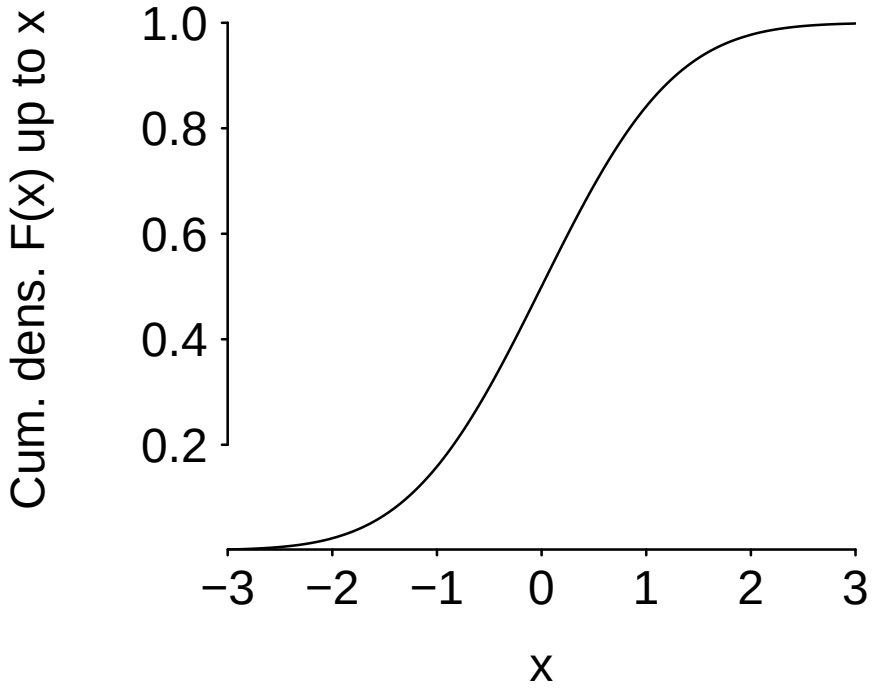
  For any possible value of $Y = y$, gives the probability that Y takes on that value

- **cdf: cumulative density function,** $F(Y)$

  For any possible value of $Y = y$,
  gives the probability that Y is less than or equal 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

- **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:

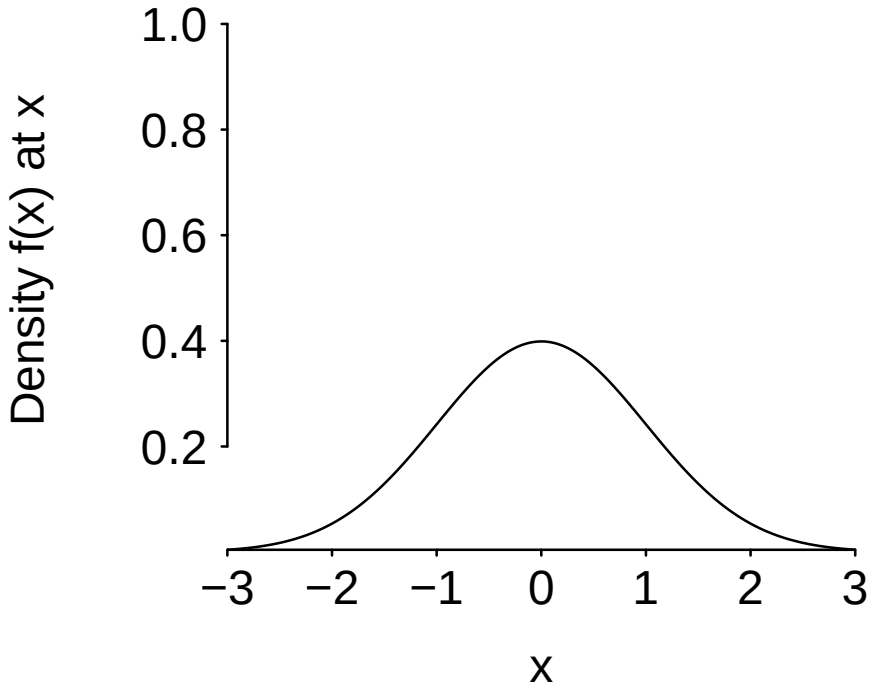
$$F(Y) = \int_{-\infty}^{y} f(Y)\mathrm{d}y$$

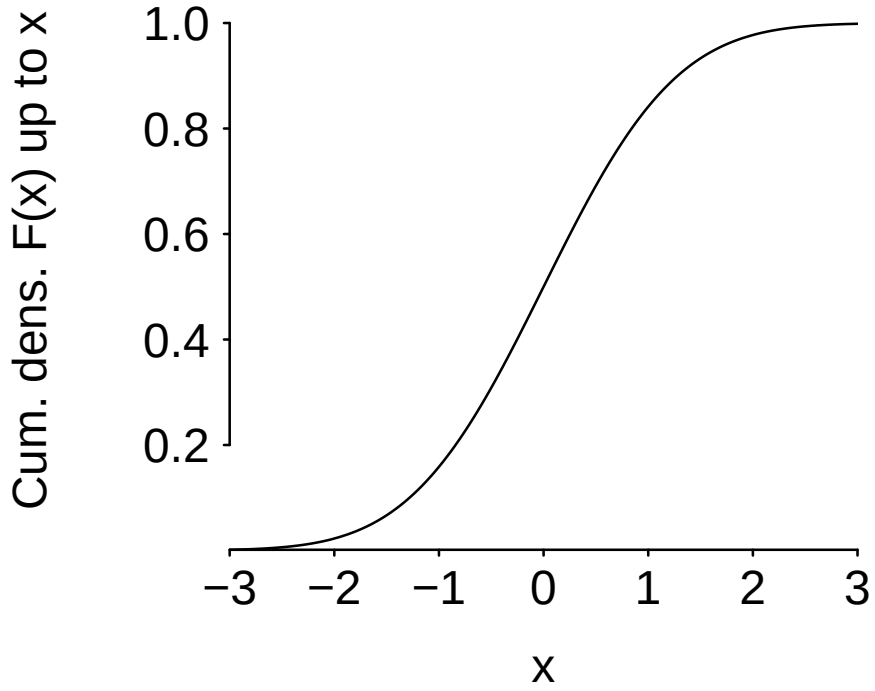
The Normal (Gaussian) distribution

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

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

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





Choosing distributions

Many researchers implicitly or explicitly assume their data is Normally distributed

But there are other distributions available

Different statistical methods are built on different probability distributions

Quantitative political scientists *select* statistical models for their data
to match the observed properties of their data

The method we'll focus on, linear regression, is based on the Normal distribution

To note right now: the Normal is a model of continuous variables, like income

So it can't be used to model discrete variables, like vote choice

Discrete data require more advanced statistical methods

Key topic in POLS/CSSS 510 (and later this quarter)

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

The Central Limit Theorem

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 $N \to \infty$

The Central Limit Theorem

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

Flipping coins

The distribution of a coin flip is $\Pr(\text{Heads}) = 0.5$, $\Pr(\text{Tails}) = 0.5$,
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.

Is it?

`http://faculty.washington.edu/cadolph/221/221lec6.pdf`
has a demonstration

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

The Central Limit Theorem

Dropping balls

`http://www.mathsisfun.com/data/quincunx.html`

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

The Normal distribution

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,
is 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:

$$\begin{array}{ccccccc} \mathbf{y} & = & \mathbf{X} & \boldsymbol{\beta} & + & \boldsymbol{\varepsilon} \\ n \times 1 & & n \times k & k \times 1 & & n \times 1 \end{array}$$

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, $\boldsymbol{\Sigma}$

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

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

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

$$\boldsymbol{\Sigma} = \operatorname{E}(\boldsymbol{\varepsilon} \boldsymbol{\varepsilon}')$$

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:

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

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

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:

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

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

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

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

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

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.

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.

What makes an estimator good?

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

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?

Three 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, a little thought shows this is not the case (diagram)

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 even be horrible.

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: 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

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:

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

$\sqrt{MSE}$ 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: 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?

We already know two things that can go wrong:

- omitted variable bias
- specification bias

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

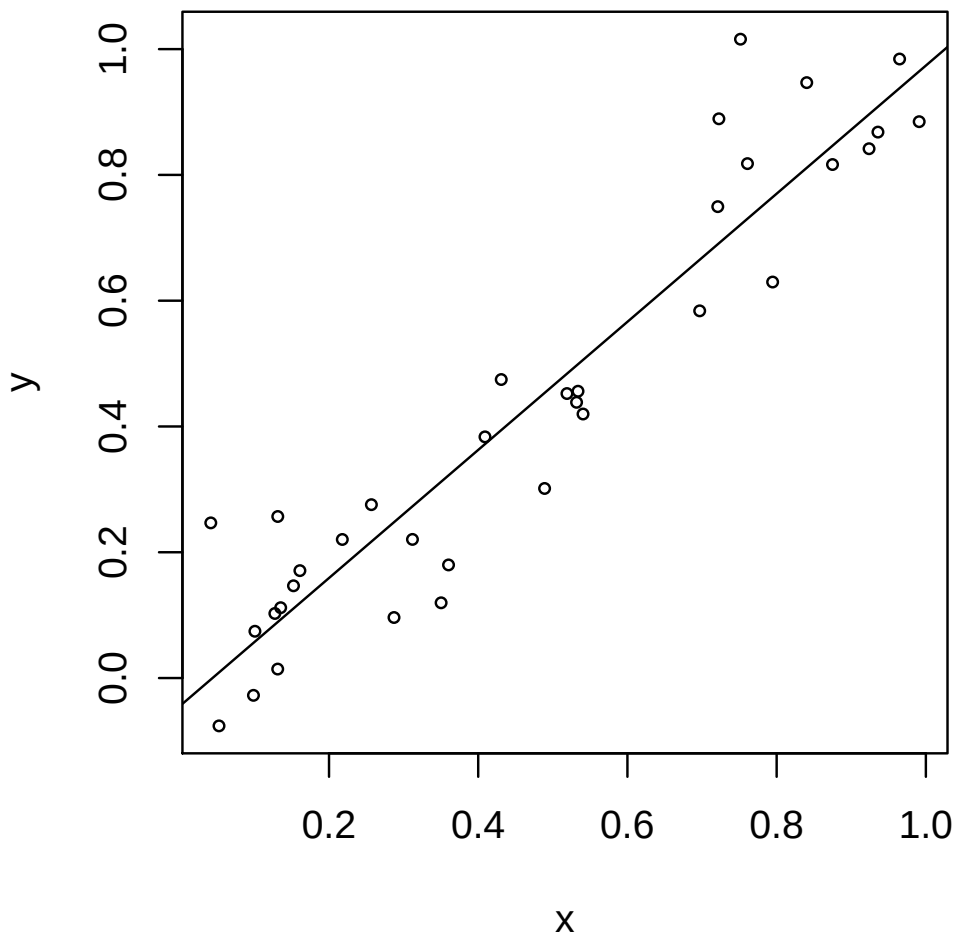
If we select observations non-randomly from the world, we may get biased estimates of means, regression coefficients, and other quantities

- 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 the Hawaii state senate,
  motorcyclists testified that in their (in some cases, multiple) crashes,
  helmets would not have prevented injuries. Who didn't testify?

- Regression example: Selection on the observed variables

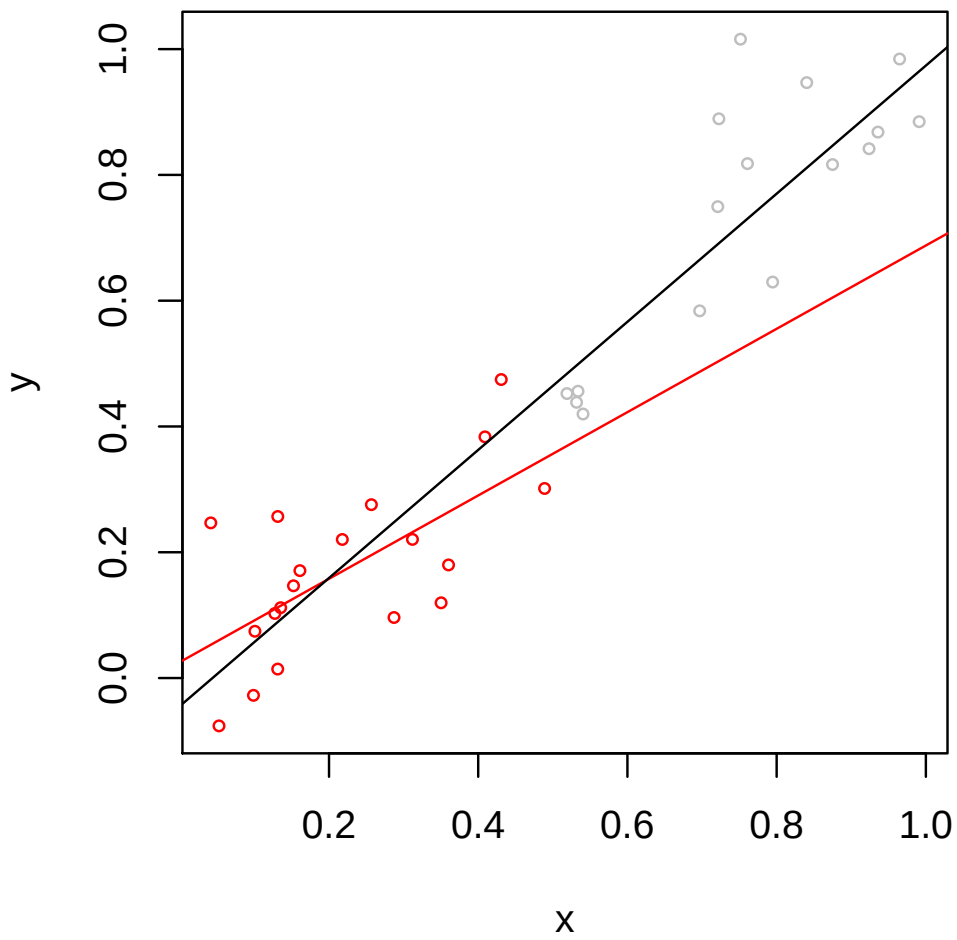
Selection bias



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

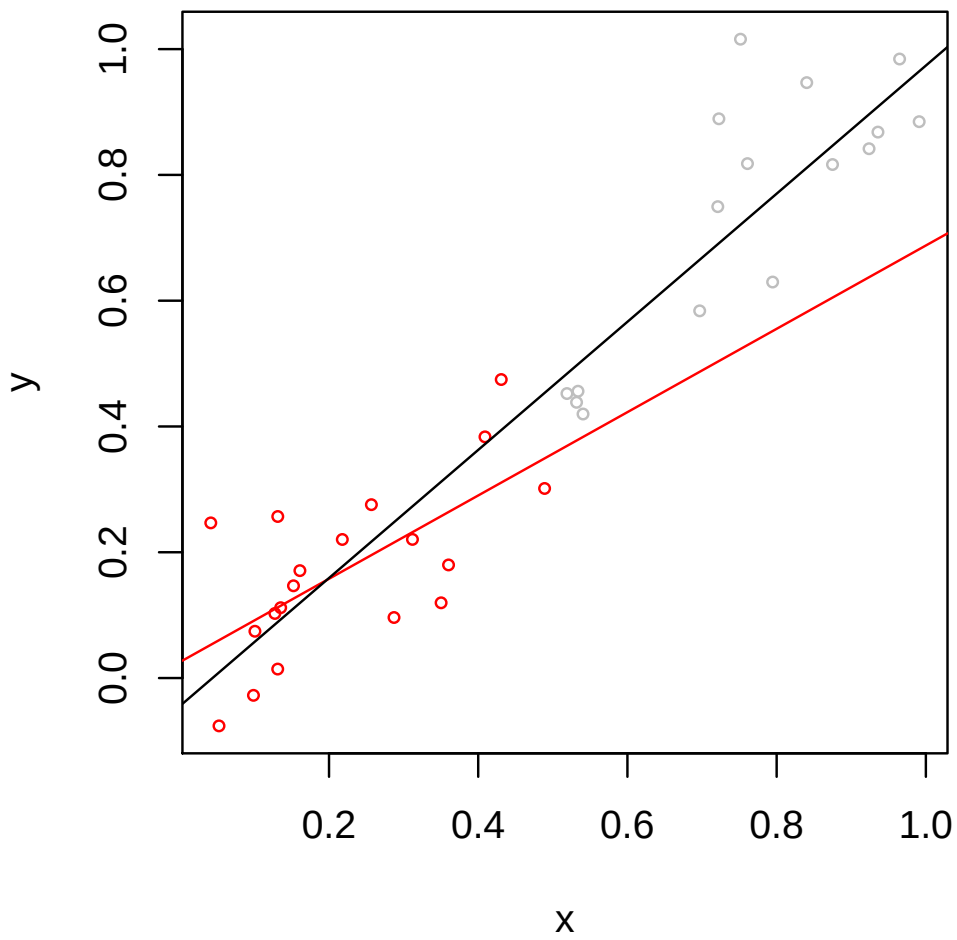


But suppose high income (or highly conservative) people decline to answer

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

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

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 under collinearity

Note an implication: if you can estimate the coefficients of all the columns of X, then there must not be any collinearity in X

If people say an estimated regression is subject to "collinearity,"
they are badly abusing terminology and likely referring to a non-problem

"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"

"Partial Collinearity" is 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's is "high"

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

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

Get more data or admit you can't distinguish between highly correlated covariates

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"

Exogenous & endogenous variables

What if $\mathbf{y}$ helps determine $\mathbf{X}$ 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
- institutions and behavior, etc.

In these cases, $\mathbf{y}$ and $\mathbf{X}$ are both endogenous

Endogeneity and causal identification

Assume it is possible that $\mathbf{y}$ and $\mathbf{X}$ are both endogenous to each other

Now suppose you wanted to **identify** the causal effect of $\mathbf{X}$ on $\mathbf{y}$ using linear regression

If $\mathbf{y}$ could also influence $\mathbf{X}$, then linear regression will *not* identify this effect

Instead, $\boldsymbol{\beta}$ is a biased estimator of the true causal effect

It will remain biased even as you add more data

In other words, it is *inconsistent*, or biased even as $N \to \infty$

Endogeneity and causal identification

Suppose taking aspirin lowers headache pain by 10 points on average
when pain is measured on a 100 point scale (true causal effect)

Let's try to recover the true causal effect from an observational study of aspirin users

We survey 1000 people and collect two variables:

Aspirin Did you take an aspirin today?

HP On a scale of 0 to 100, how much did your head hurt today?

We estimate by linear regression:

$$\mathrm{HP}_i = \beta_0 + \beta_1 \mathrm{Aspirin}_i + \varepsilon_i$$

and expect to find $\hat{\beta}_1 = 10$. But instead, we find $\hat{\beta}_1 = -60$! Why?

Having headache pain makes people much more likely to take aspirin,
which is only partly effective at mitigating that pain

the covariate (Aspirin) is endogeous to the outcome (HP)!

Endogeneity and causal identification

How could you identify the causal effect of $\mathbf{X}$ on $\mathbf{y}$?

Strong research design, usually using one of these techniques:

1. Controlled experiments (in the field and lab)
2. Instrumental variables regression
3. Regression discontinuity designs
4. "Natural" experiments, often combined with matching or regression

Note that controlled experiments tend to be more strongly identified that the other options, but may have limited external validity

More on causal identification later in the course;
now, back to problems and properties of linear regression

Heteroskedasticity: "Different variance"

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

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

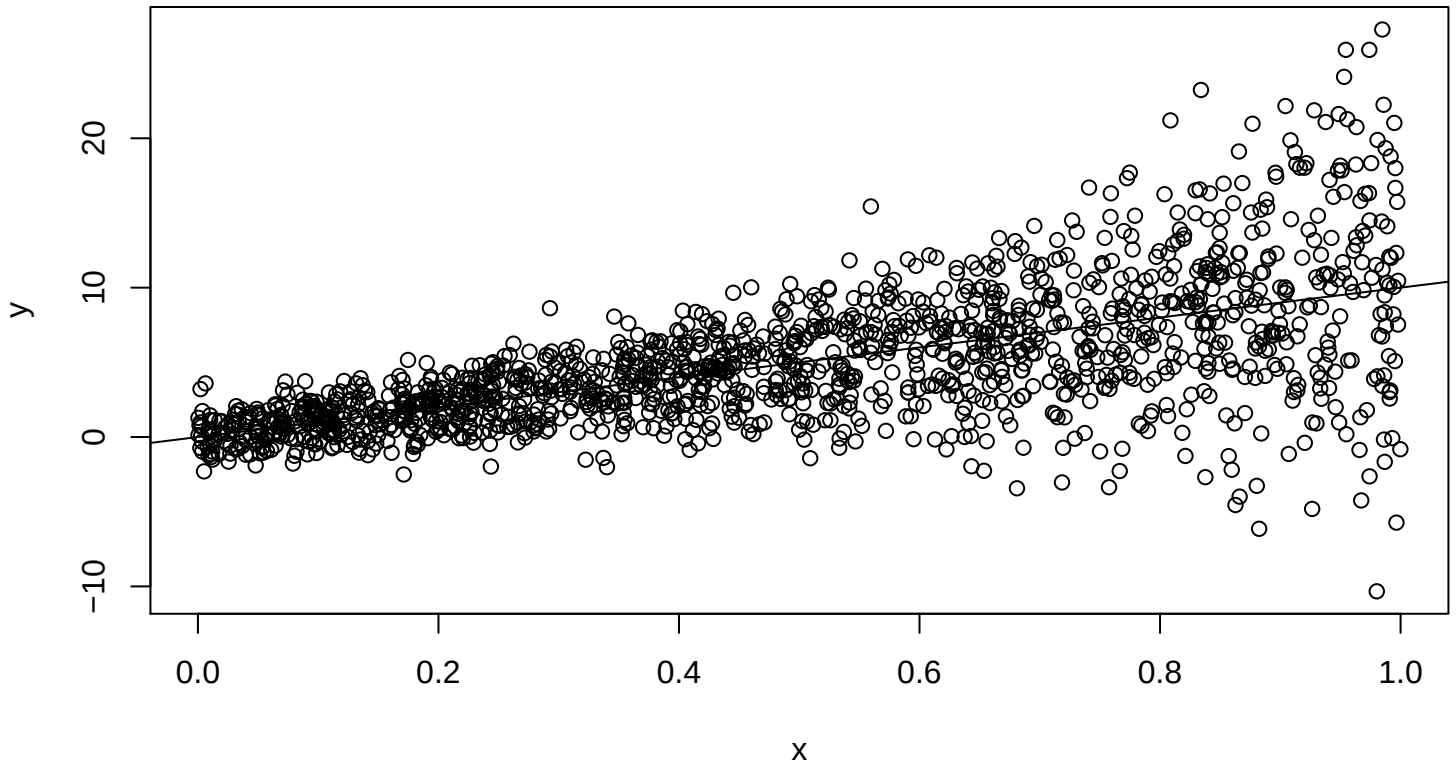
But LS always assumes the variance of that variable is the same:

σ^2, a constant

We don't think $\mathbf{y}$ has constant mean. So why expect constant variance?

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

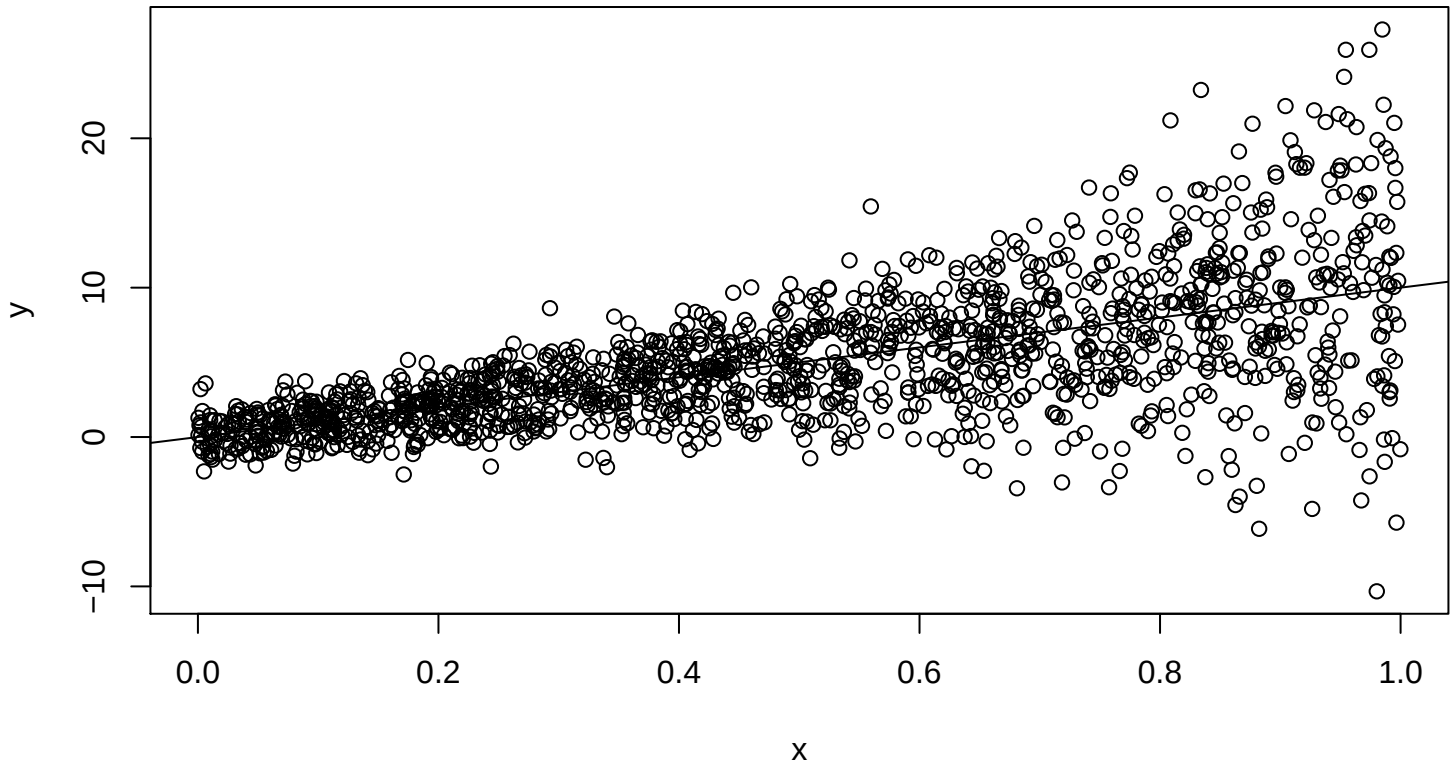
Heteroskedasticity: "Different variance"



A common pattern of heteroskedasticity:
Variance and mean increase together

Here, they are both correlated with the covariate x

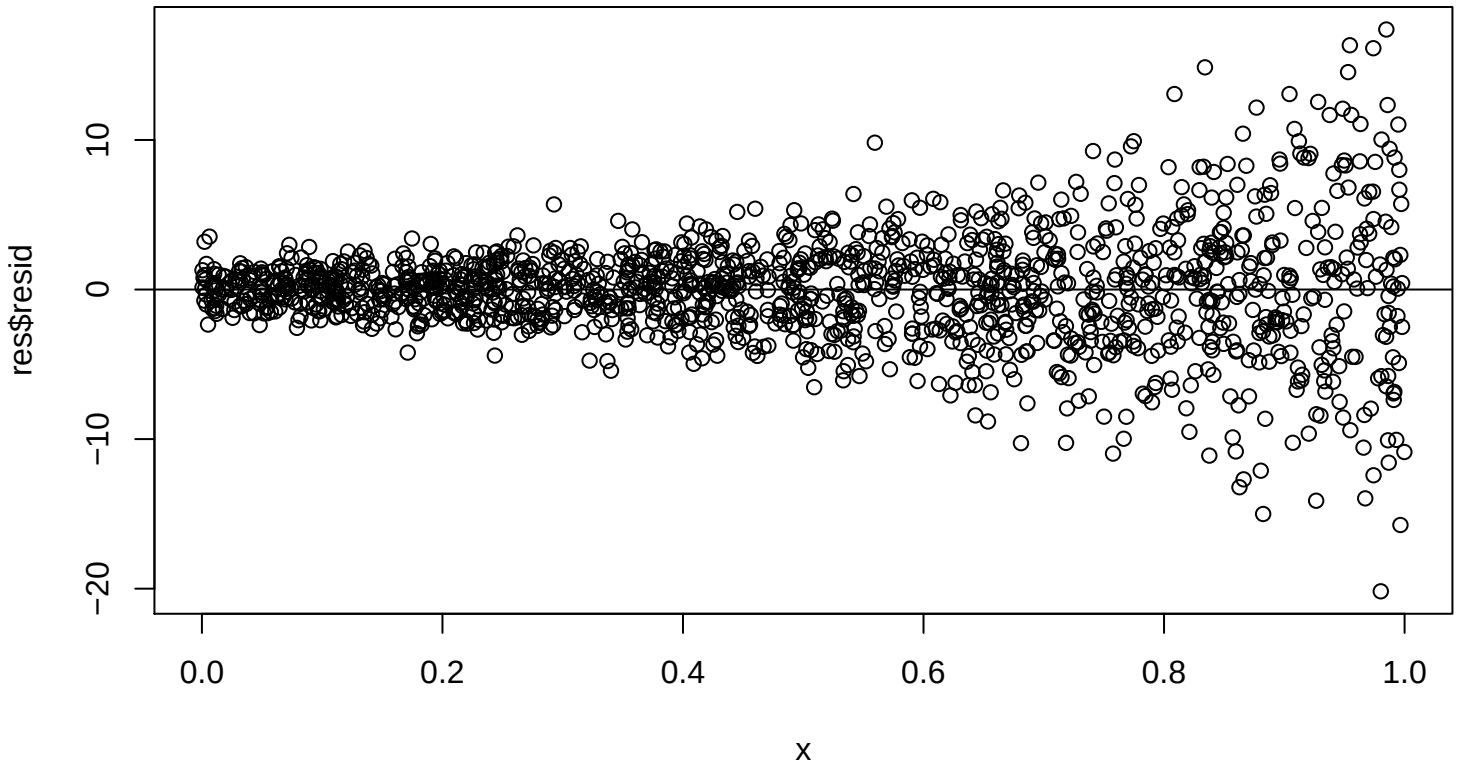
Heteroskedasticity: "Different variance"



In a fuzzy sense, x is a necessary but not sufficient condition for y

This is usually an important point substantively. Heteroskedasticity is *interesting*, not just a nuisance

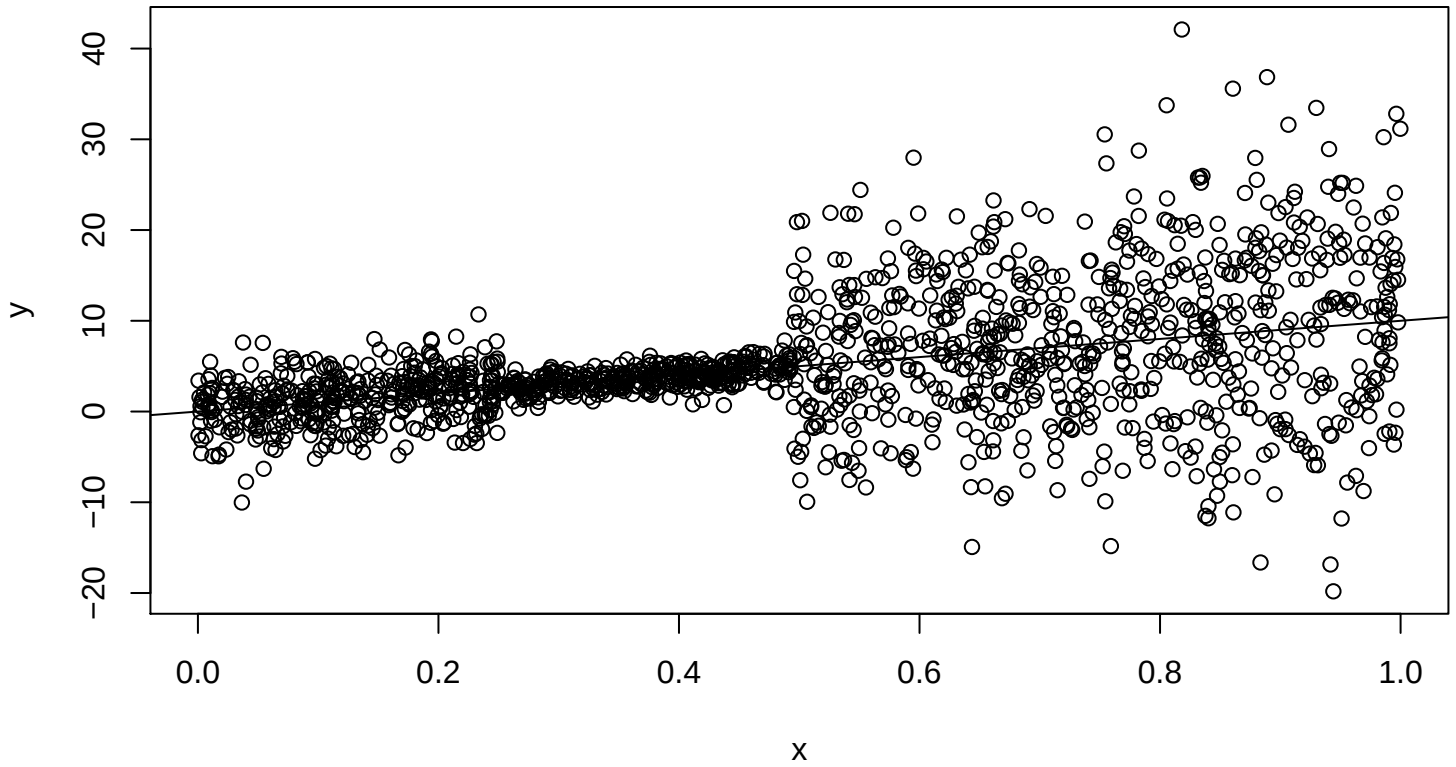
Heteroskedasticity: "Different variance"



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

Look for a pattern in the *variance* (different from a pattern in the mean).

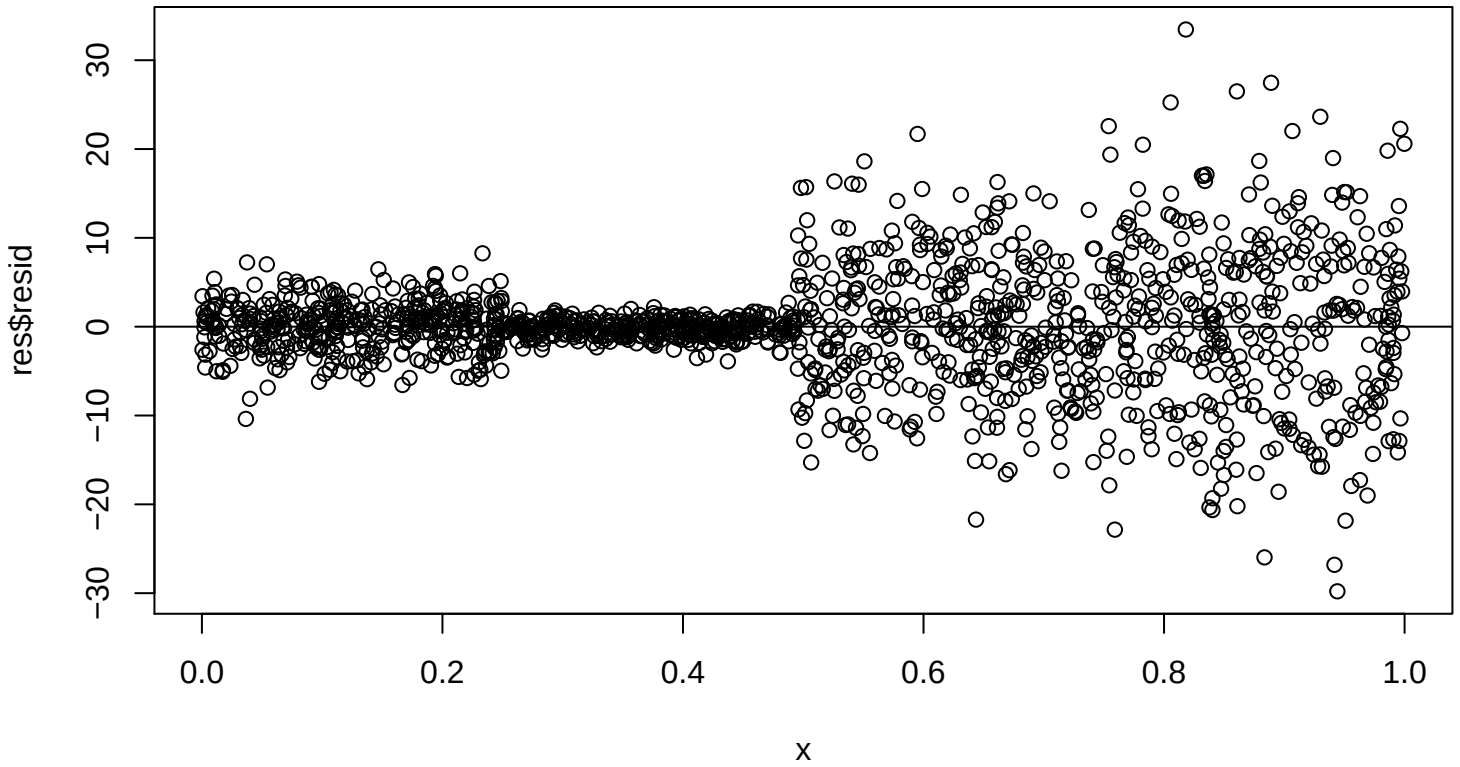
Heteroskedasticity: "Different variance"



Often a megaphone shape, 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. . .

The Multivariate Normal distribution

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

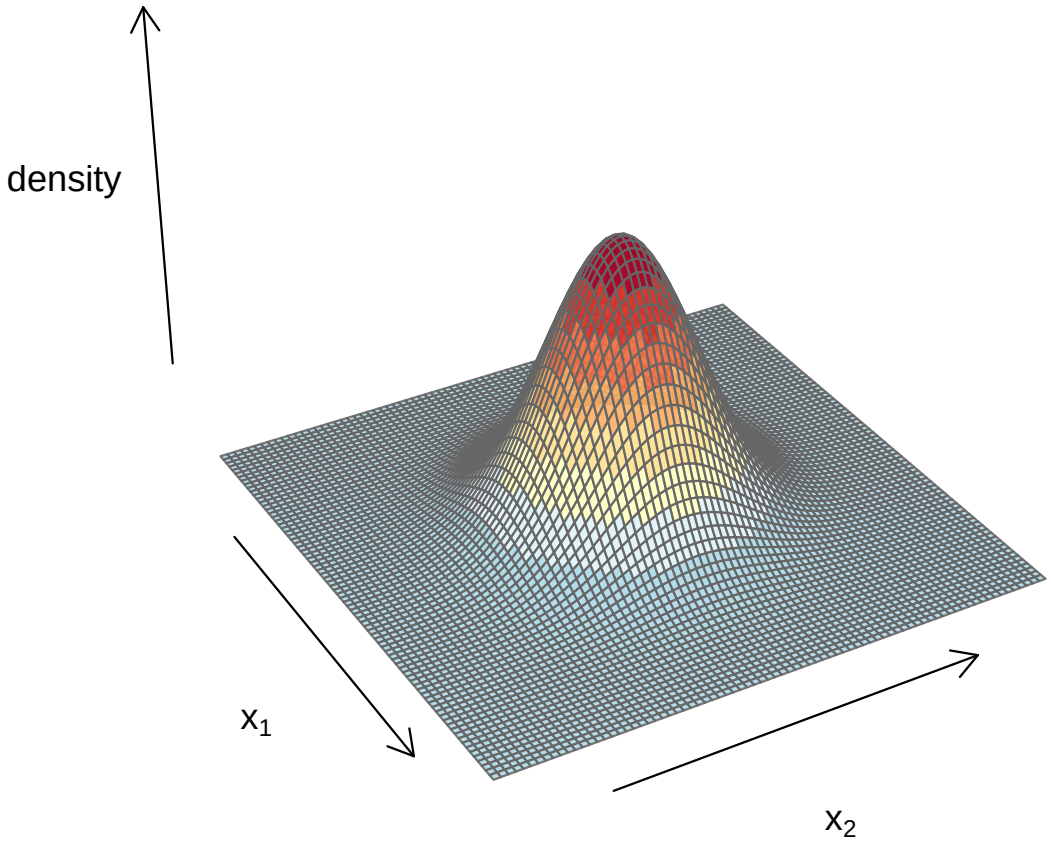
As usual, let μ indicate a mean, and σ a variance or covariance

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

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

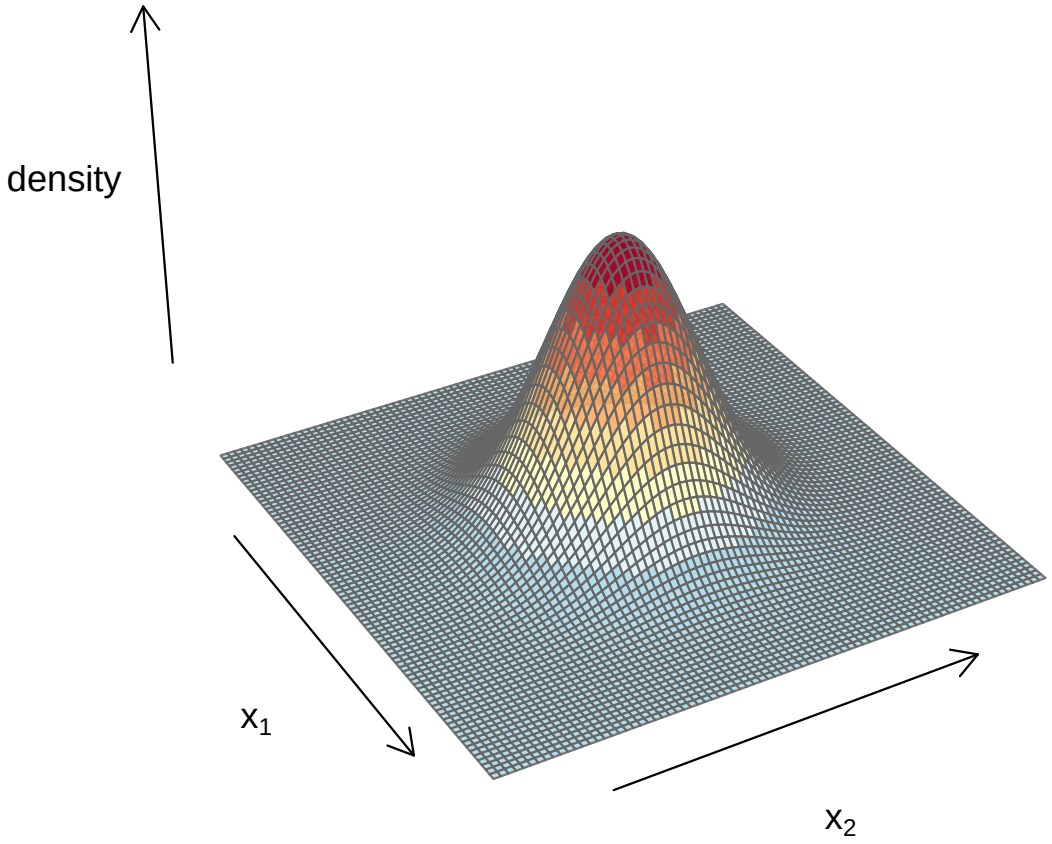
The MVN is more than the sum of its parts:
There is a mean and variance for each variable, *and* covariance between each pair

The Multivariate Normal distribution



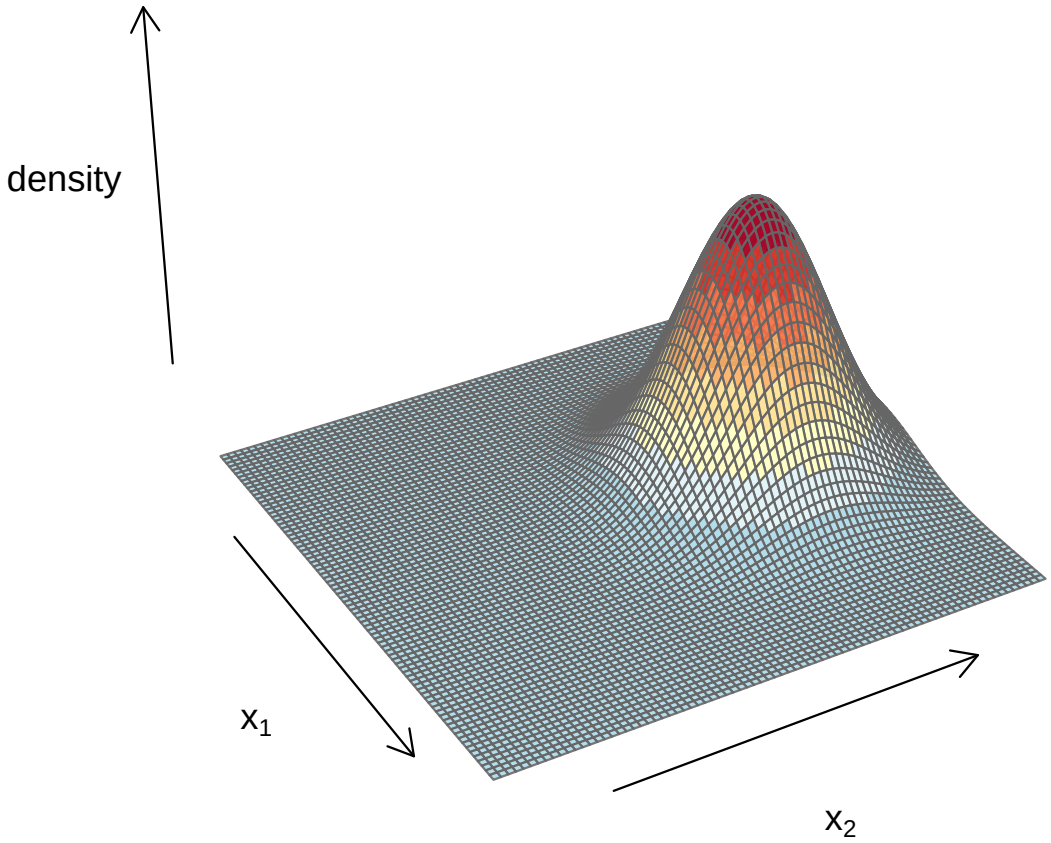
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



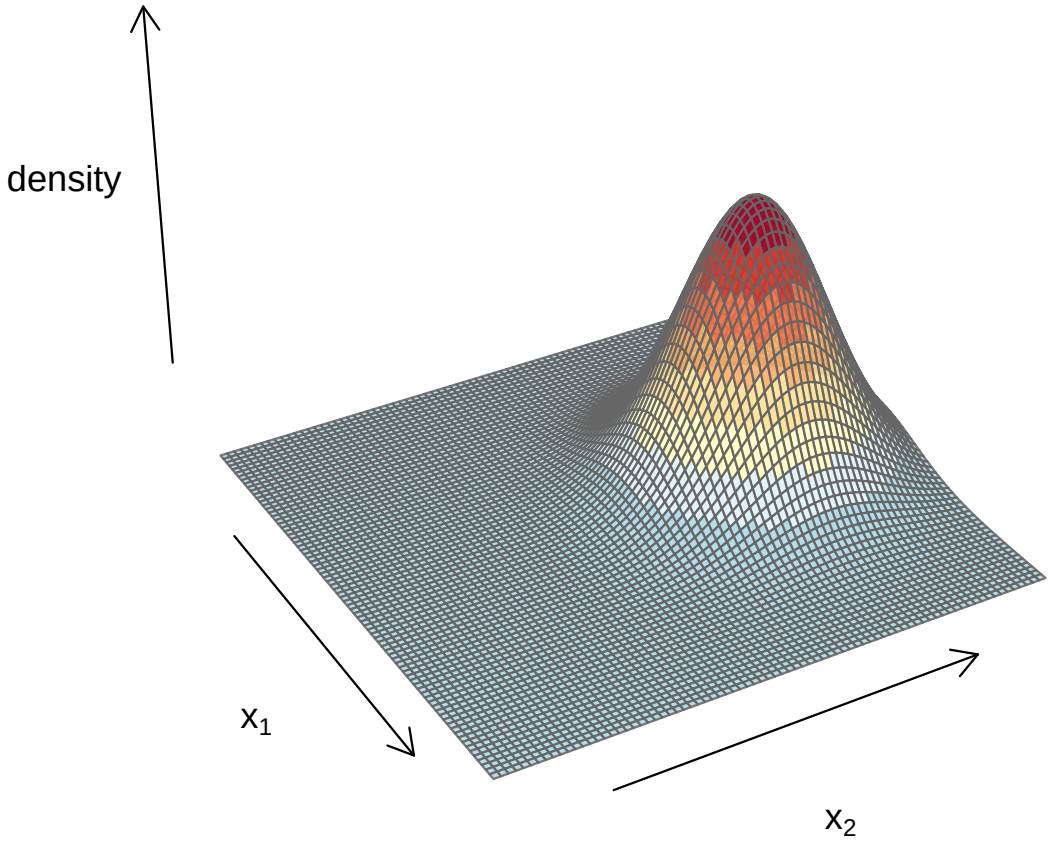
The standard MVN, with zero means, unit variances, and no covariance, looks like a higher dimension version of the normal: a symmetric mountain of probability

The Multivariate Normal distribution



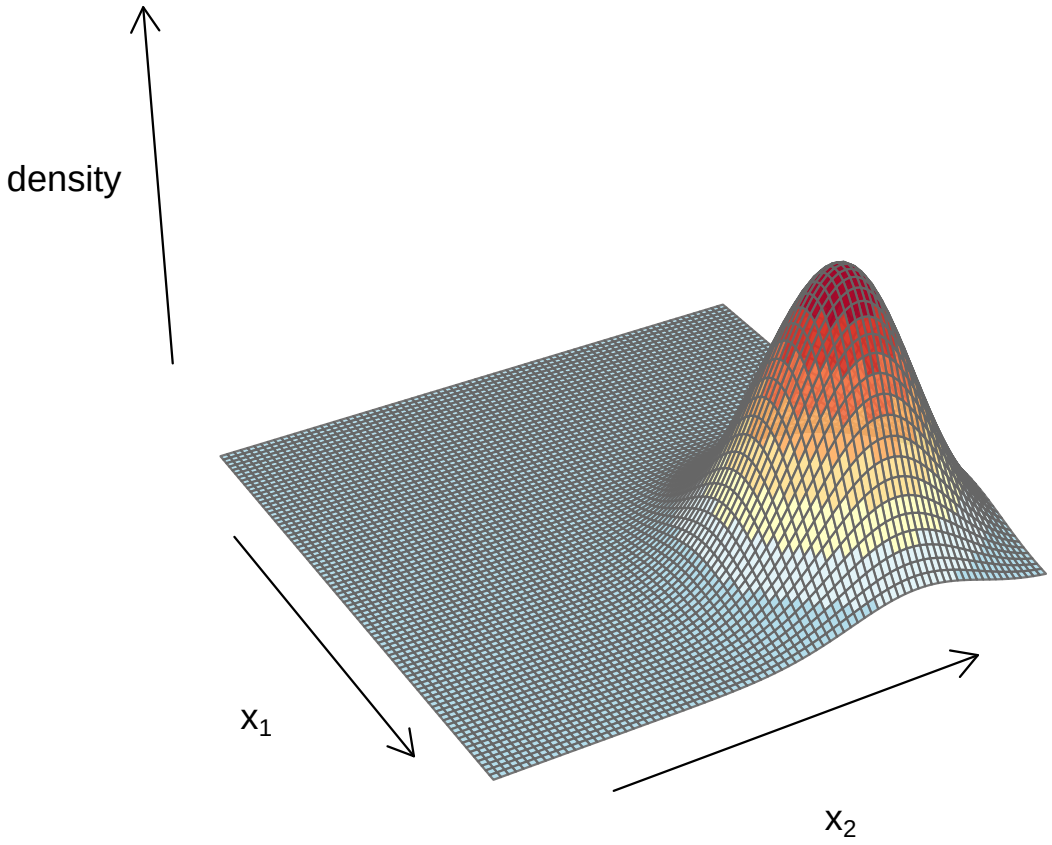
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



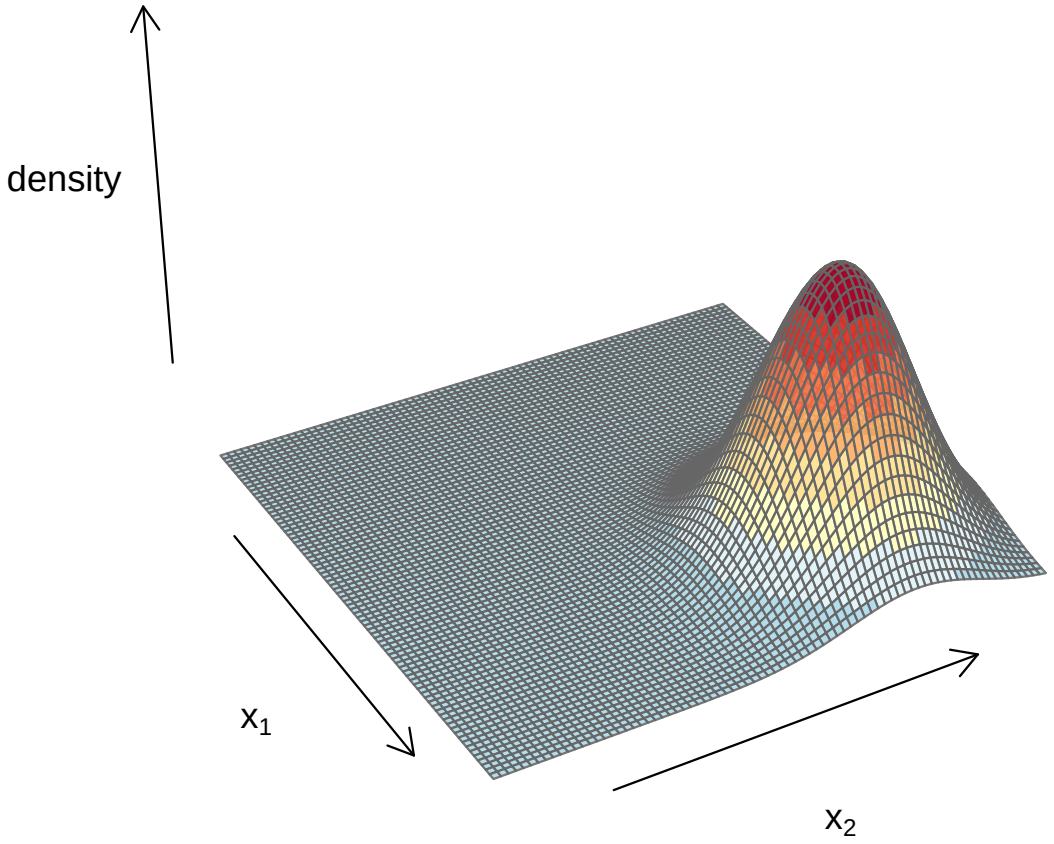
Shifting the mean of $\mathbf{x}_2$ moves the MVN in one dimension only
Mean shifts affect only one dimension at a time

The Multivariate Normal distribution



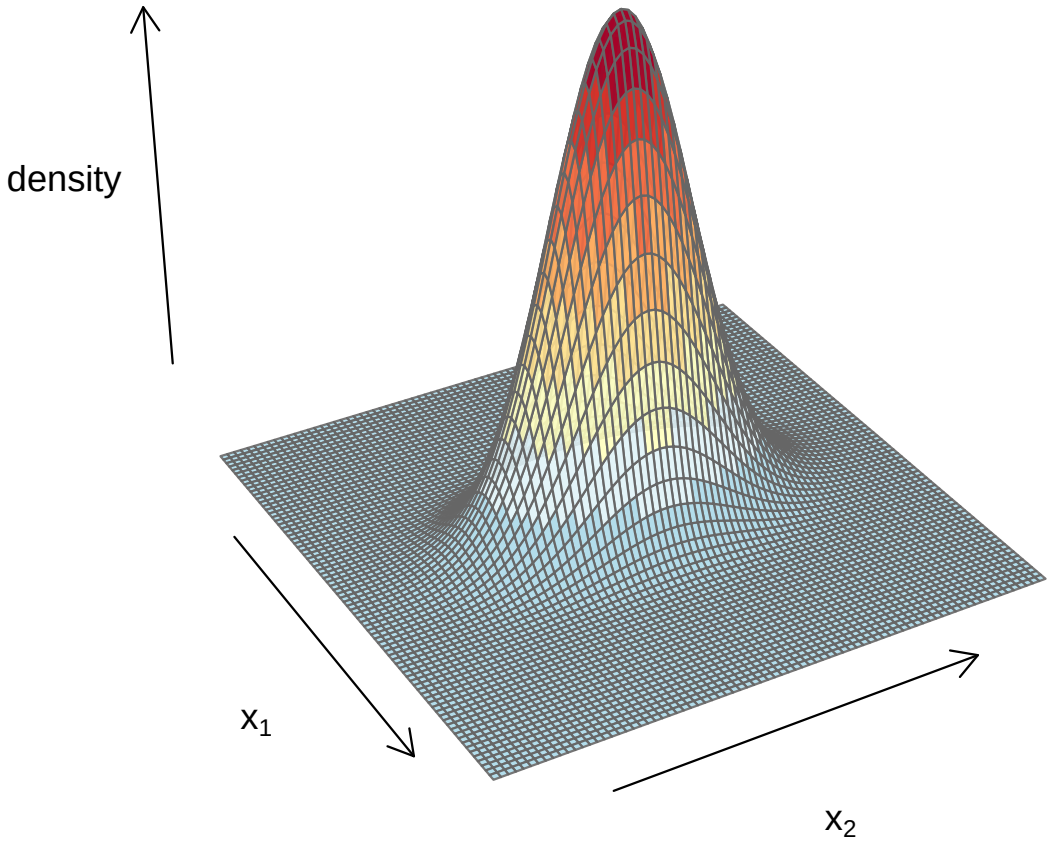
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



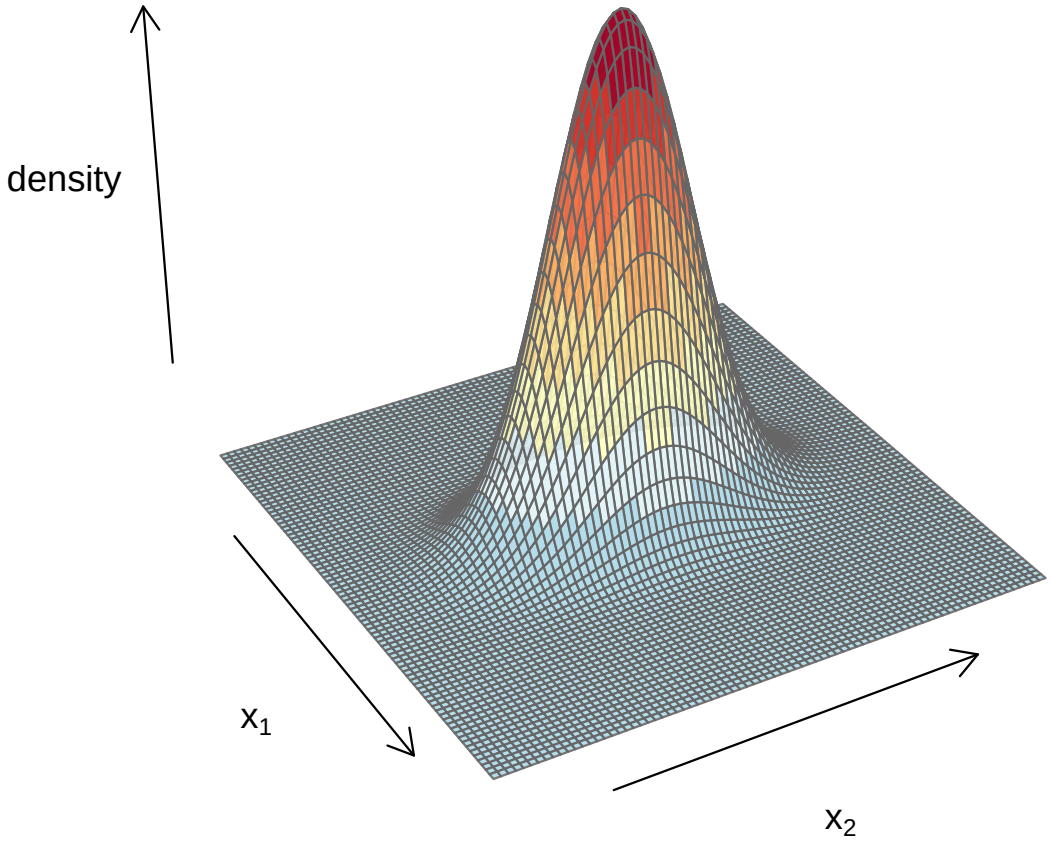
We could, of course, move the means of our variables at the same time
This MVN says the most likely outcome is both x_1 and x_2 will be near 2.0

The Multivariate Normal distribution



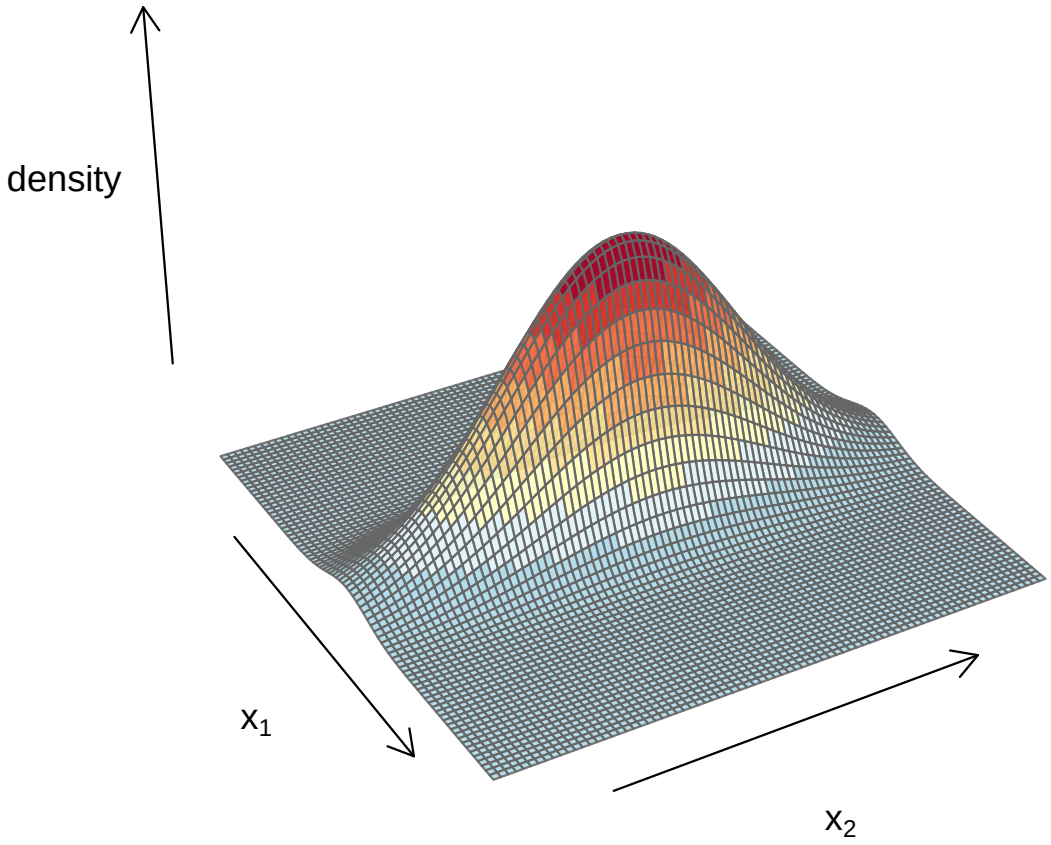
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



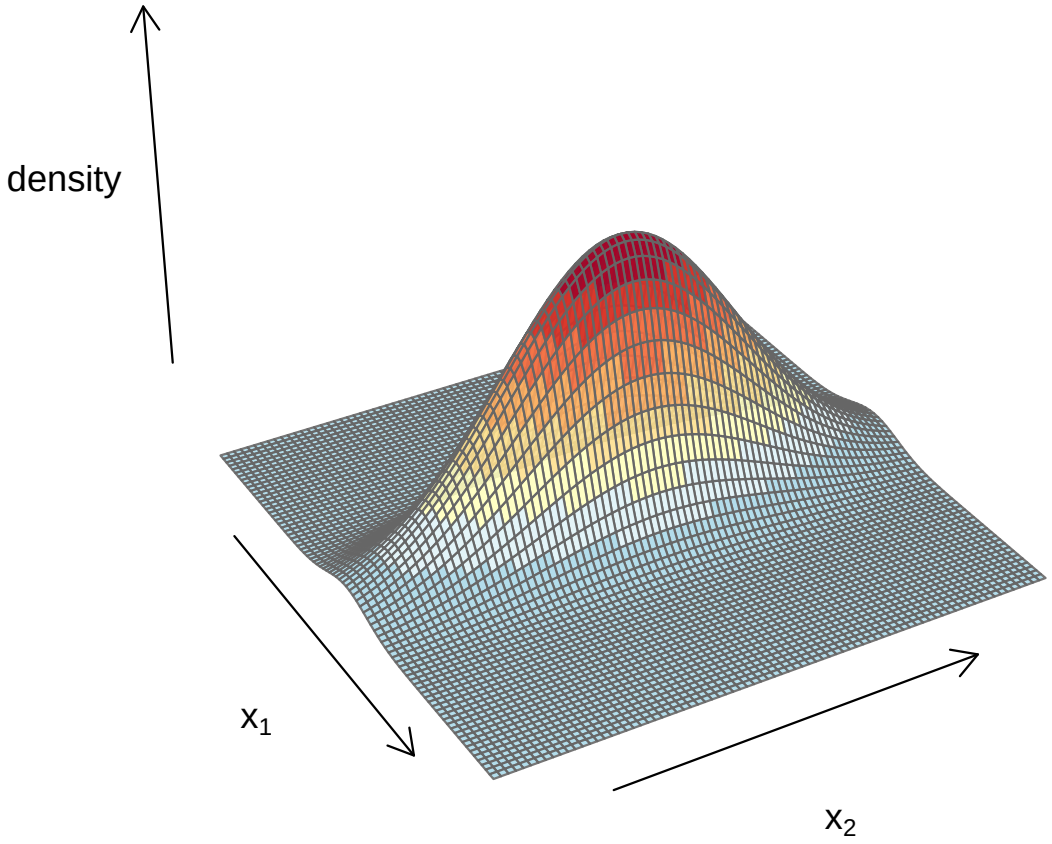
Shrinking the variance of $\mathbf{x}_1$ moves the mass of probability towards the mean of $\mathbf{x}_1$, but leaves the distribution around x_2 untouched

The Multivariate Normal distribution



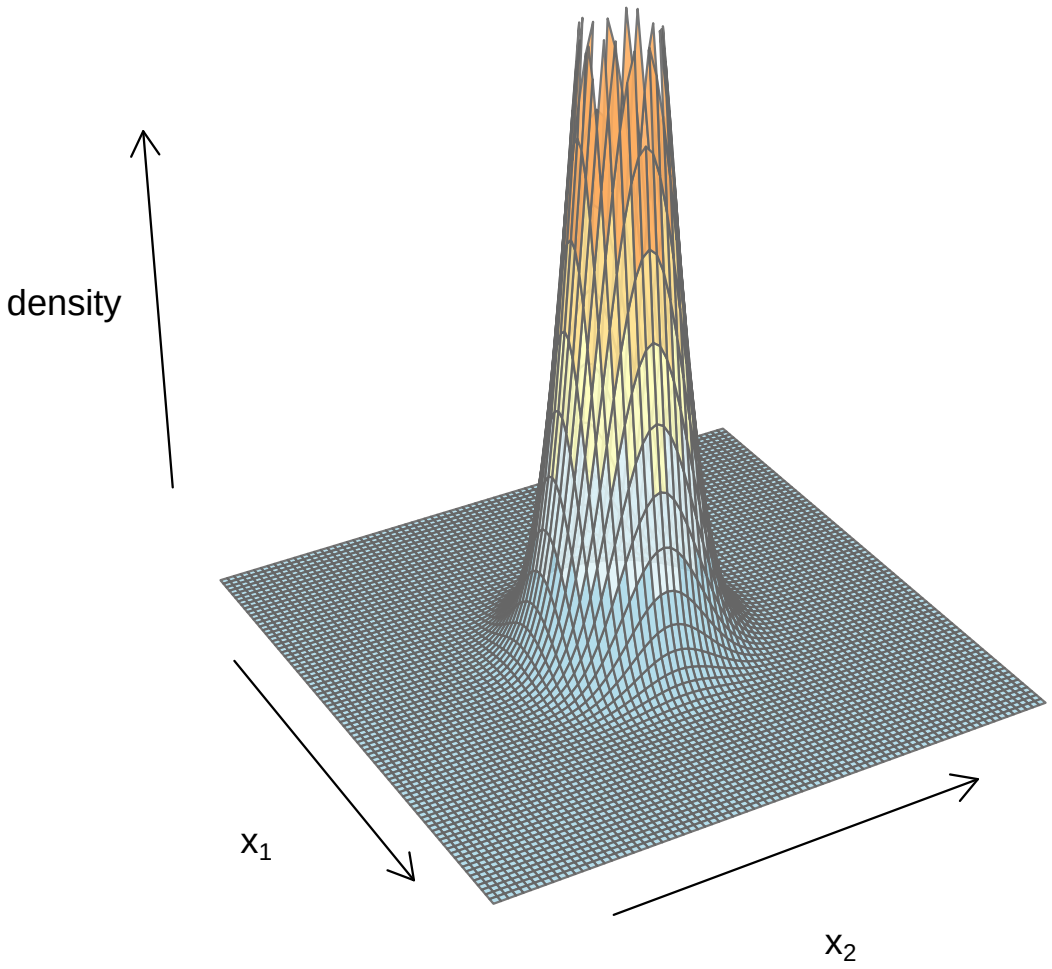
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 3 \end{bmatrix} \right)$$

The Multivariate Normal distribution



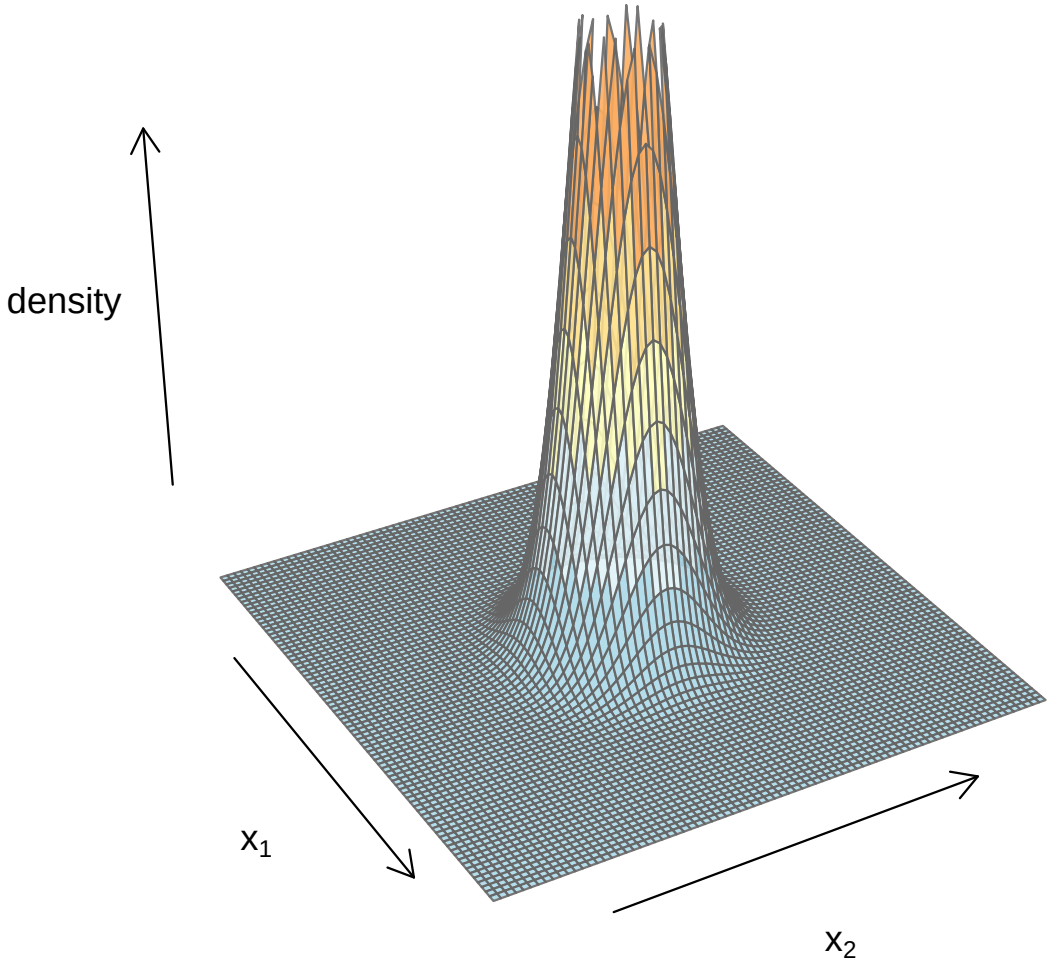
Increasing the variance of $\mathbf{x_2}$ spreads the probability out,
so we are less certain of $\mathbf{x_2}$, but just as certain of $\mathbf{x_1}$ as before

The Multivariate Normal distribution



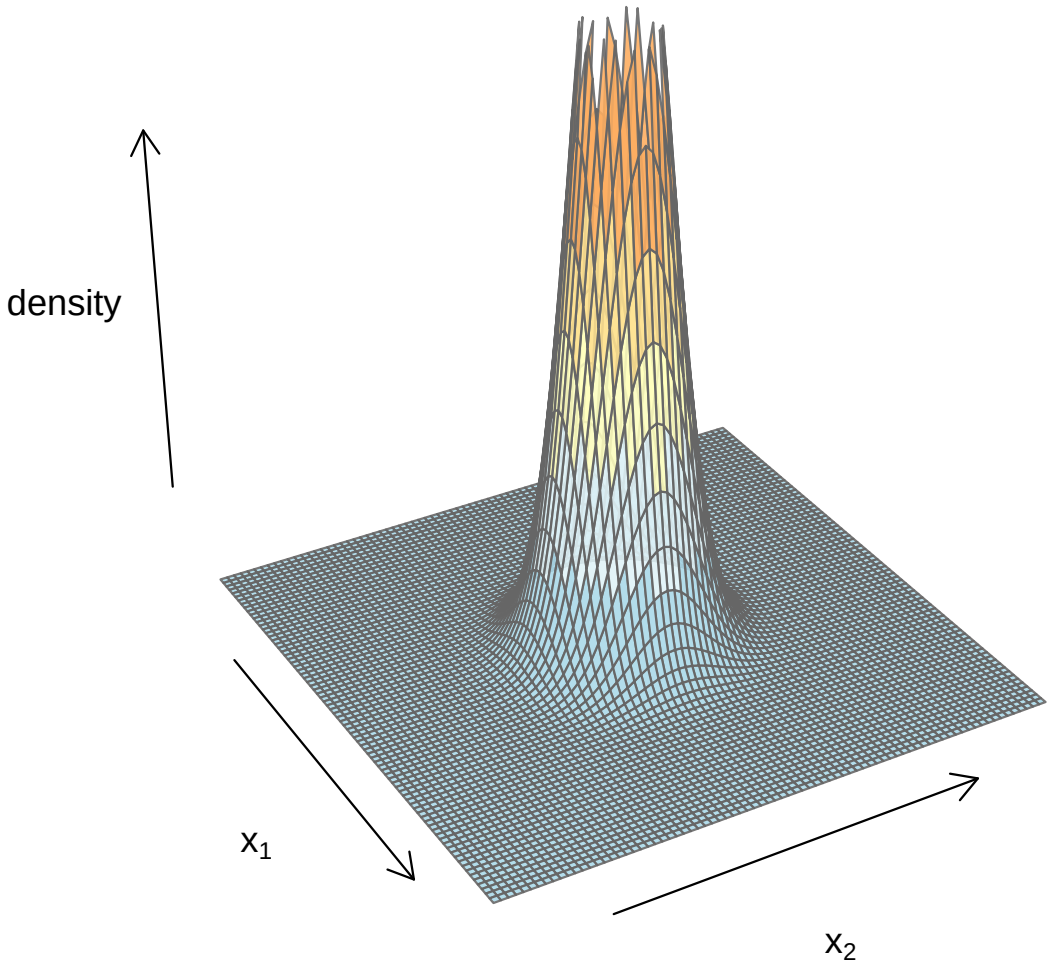
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 0.33 \end{bmatrix} \right)$$

The Multivariate Normal distribution



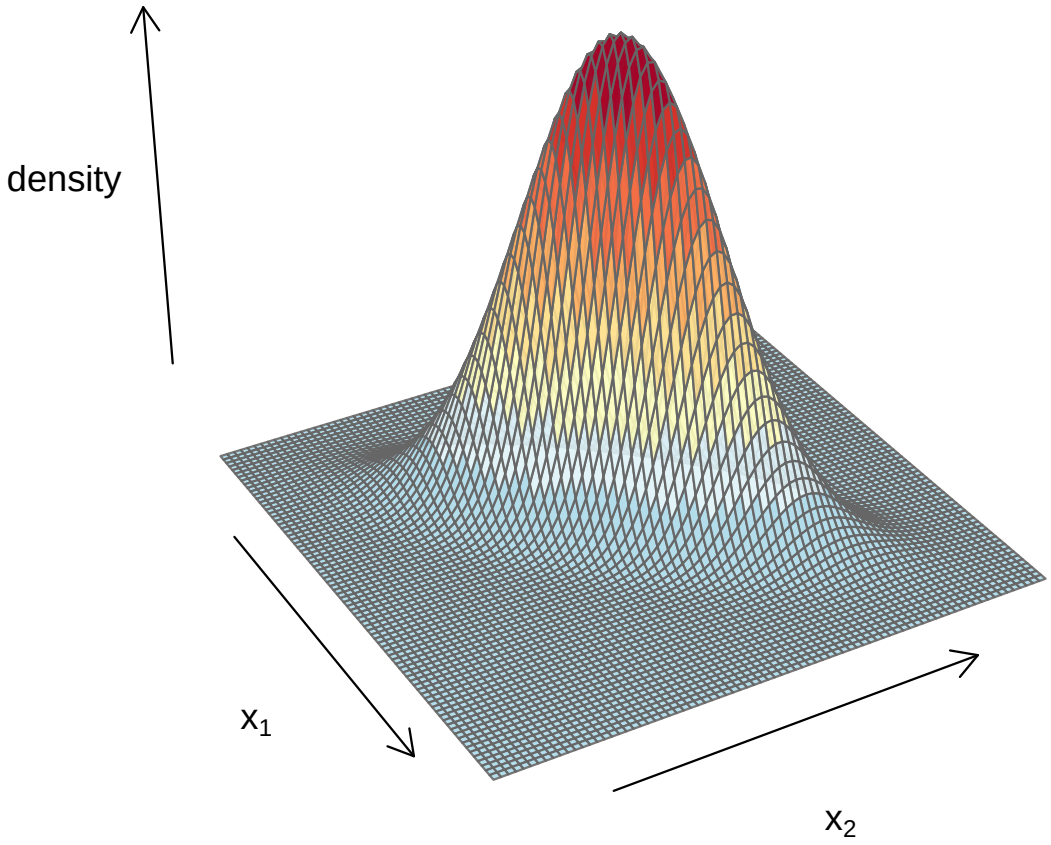
If the variance is small on all dimensions,
the distribution collapses to a spike over the means of all variables

The Multivariate Normal distribution



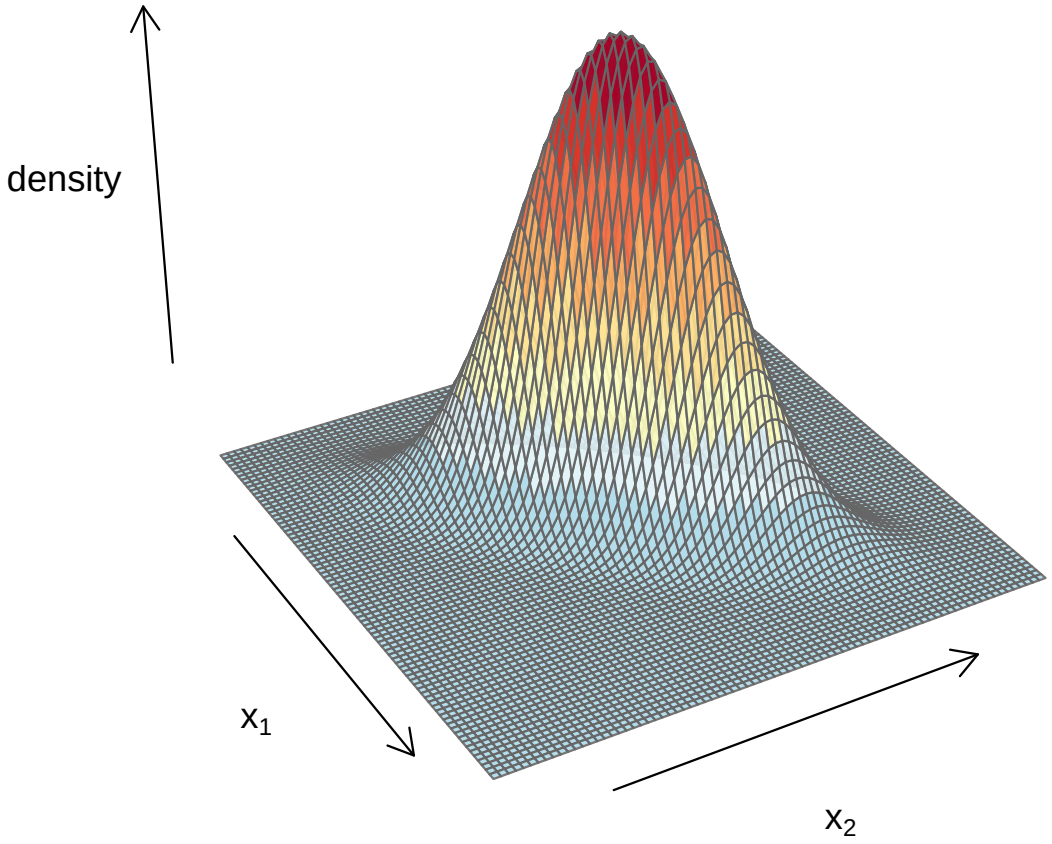
In this case, we are fairly certain of where all our variables tend to lie

The Multivariate Normal distribution



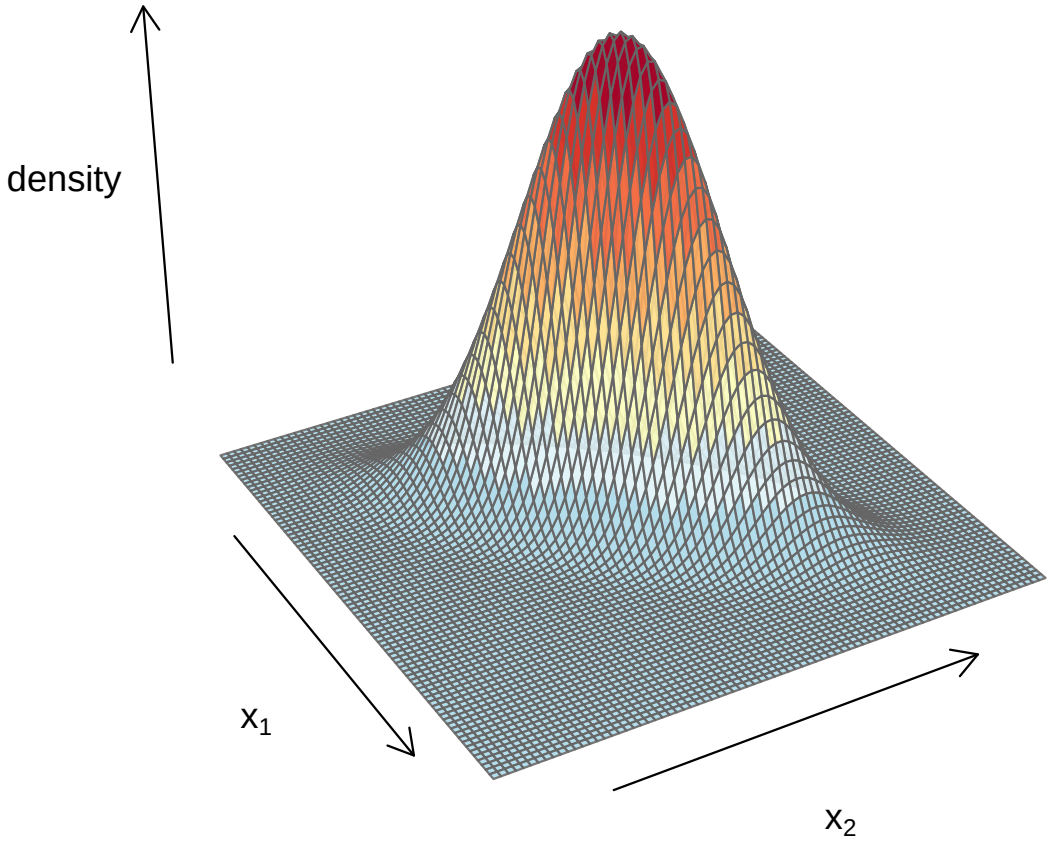
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}\right)$$

The Multivariate Normal distribution



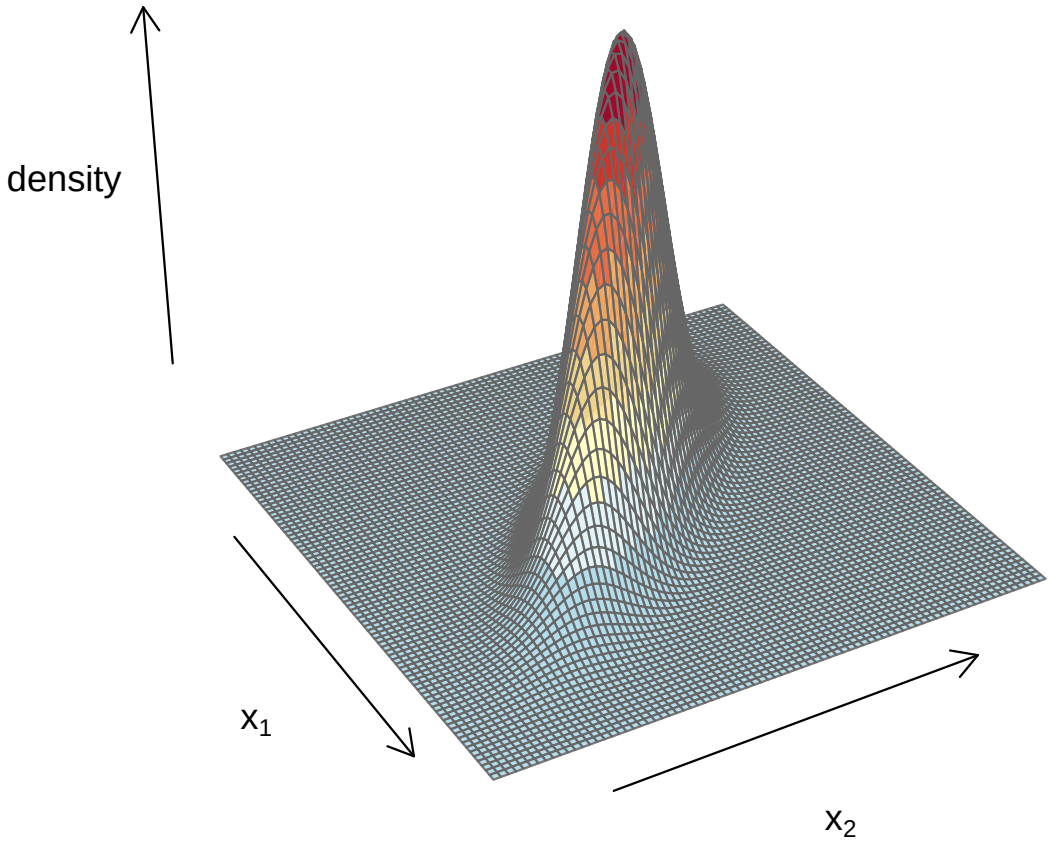
In this special case, with unit variances, the covariance is also the correlation, so our distribution say x_1 and x_2 are correlated at $r = 0.8$

The Multivariate Normal distribution



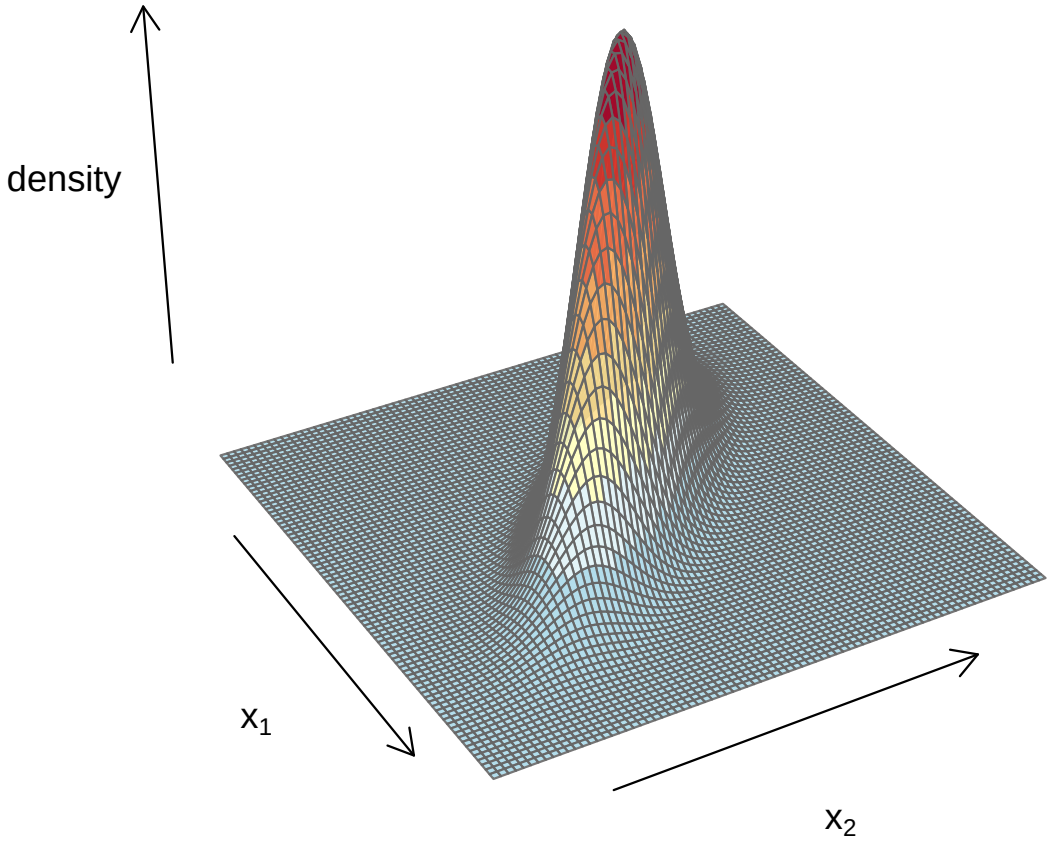
A positive correlation between our variables makes the MVN asymmetric, with greater mass on likely combinations

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

The Multivariate Normal distribution

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

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

Unpacking σ^2: homoskedastic case

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

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

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

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 in many $n + 1$ dimensions

We say that errors are "spherical" when this symmetry holds

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

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

If we fail to capture this dynamic, our errors will be correlated

(Here, MVN is the "tilted" ridge: any shape is possible)

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
1	No (perfect) collinearity	$\text{rank}(\mathbf{X}) = k, k < n$	

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

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

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

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

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

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

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

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

Gauss-Markov Theorem

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 $\boldsymbol{\varepsilon}$, and
- $E(\boldsymbol{\varepsilon}) = 0$ (regardless of the distribution of $\boldsymbol{\varepsilon}$)

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)

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

R programming

So far, our R programs have had a simple structure

We have simple told R to do one thing after another

But we can write much more powerful programs by adding structure:

`{}`	Grouping commands
`if(cond) {} else {}`	Running either one block of code or another depending on a condition
`for(i in a:b) {}`	Looping over groups of command
`while(cond) {}`	Running a series of commands until a condition is fulfilled
`myfunction <- function(myinput){}`	Writing our own R modules

More complex structures

To group a set of R statements together, use {}; e.g.,

```
{
  x <- "hello world"
  print(x)
}
```

This comes in handy for conditional branching:

```
if (median==1) {
  y <- median(x)
} else {
  y <- mean(x)
}
```

Loops

R has two kinds of loops, `for` and `while`

```
x <- rep(1,100)    # Replicate 100 1s into a vector
x <- rep(NA,100)   # Replicate 100 NAs into a vector
for (i in 1:length(x)) {
  y[i] <- x[i] + i
}
```

`for` loops have set termination and an index;
`while` loops are more flexible

```
x <- NULL
done <- FALSE
while(!done) {
  x <- c(x,incr)
  incr <- incr + 1
  if (incr>10)
    done <- TRUE
}
```

Be careful to avoid infinite loops

Functions

The best programs consist of generalized modules,
each of solves a particular problem regardless of details.

`lm()` is an example:
this function always produces regression results,
regardless of the number of observations or variables

Novice programmers produce long blocks of code which only works for one example
e.g., regression code that only works for a specific dataset and, say, 3 covariates

Strive to be general and modular.

- including very few "numbers" in your code (use variables instead)
- write *functions* to solve problems once and for all

Functions

You can write R functions, and include them in your code (or in a separate file)

An R function looks like this:

```
name <- function(argument list) {
  expressions
  return value
}
```

for example,

```
# A factorial function
fact <- function(x){
          if (x <= 1) {
            return(1)
          }
          f <- 1
          for (i in 1:x)
            f <- f * i
          f
        }
```

Quick intro to R: Probability distributions

Some probability distributions you can get in R:

```
norm binom beta cauchy chisq exp f
gamma geom hyper lnorm logis nbinom
t unif weibull wilcox
```

These names by themselves aren't commands; you need to prefix one of four letters:

`d`	density (pdf)	`dnorm()`
`p`	distribtion (cdf)	`pgamma()`
`q`	quantile (e.g., percentiles)	`qt()`
`r`	random number generator	`runif()`

You can find more distributions (and lots more features) in add-on packages.

To load one, like MASS, type `library(MASS)` at the start of your code/session