Foundations of Linear Regression

3. Properties and Assumptions

GOVT 6029 - Spring 2021

Cornell University

What makes an estimator good?

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

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 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?

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?

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?

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?

Some common criteria:

What makes an estimator good?

Some common criteria:

Bias

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.

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**, $\mathrm{E}(\hat{\beta} - \beta)$

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**, $\mathrm{E}(\hat{\beta} - \beta)$

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.

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.

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:

What makes an estimator good?

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

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

What makes an estimator good?

Efficiency: Efficient estimators get closest to the truth on average

Some examples:

What makes an estimator good?

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		

What makes an estimator good?

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**:

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

$\sqrt{MSE}$ 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:

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

What can go wrong in the linear model?

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

What can go wrong in the linear model?

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?

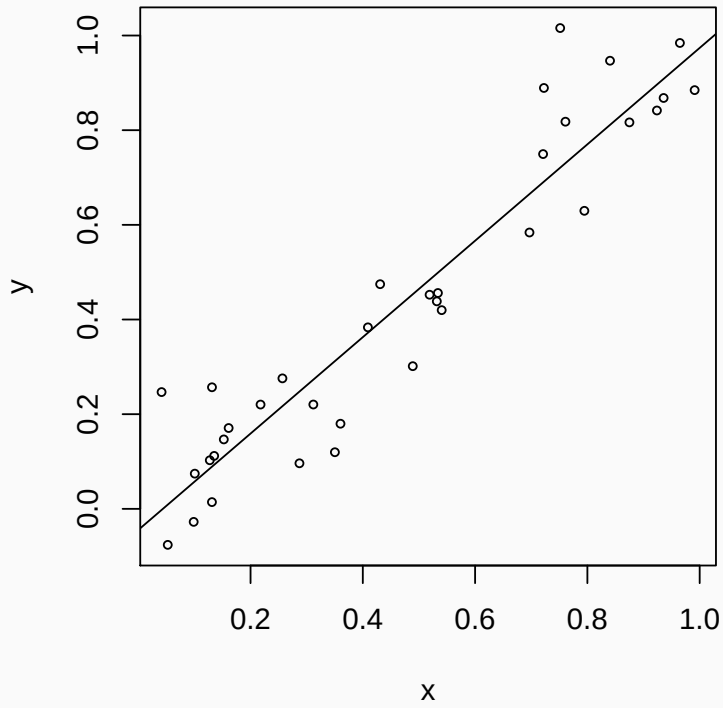
What can go wrong in the linear model?

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

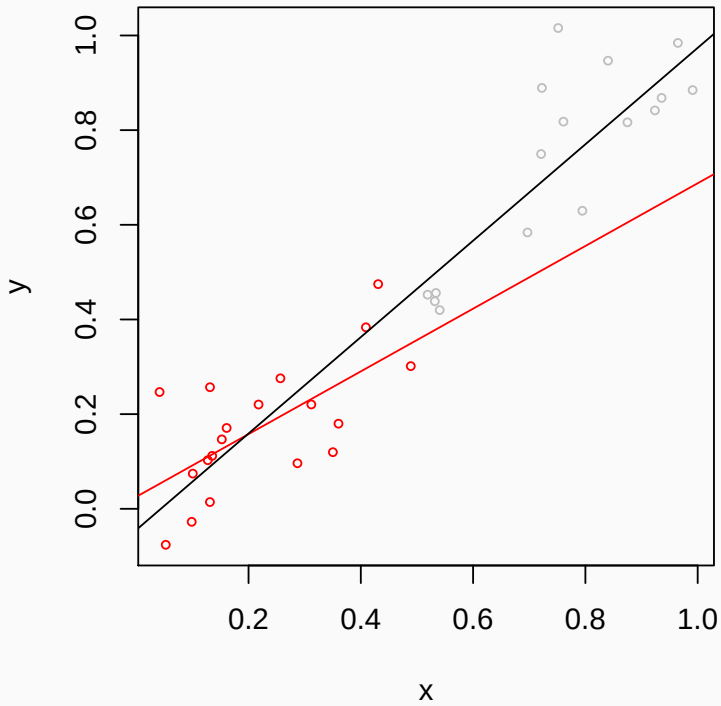
Selection bias



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

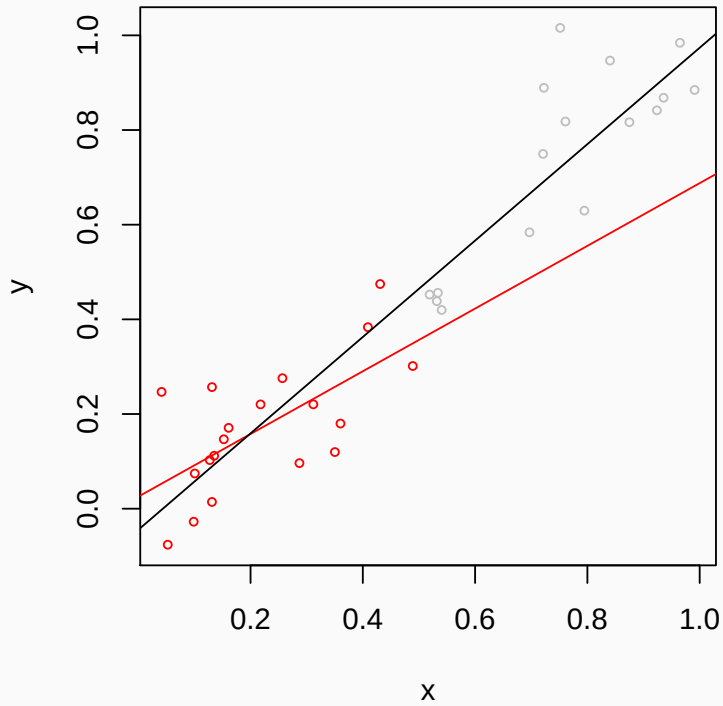


But suppose high income (or highly conservative) people decline to answer

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 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

“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 actually an oxymoron.

Inappropriately, this situation is sometimes called “multicollinearity”

“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 $\mathbf{X}$ 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 $\mathbf{X}$ is “high”.

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

Have highly correlated $\mathbf{X}$ 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 $\mathbf{X}$ is “high”.

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

Have highly correlated $\mathbf{X}$ and large se’s?
Then you lack sufficient data to precisely answer your research question

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

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 **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 inconsistent, or biased even as $N \to \infty$

Heteroskedasticity: “Different variance”

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

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

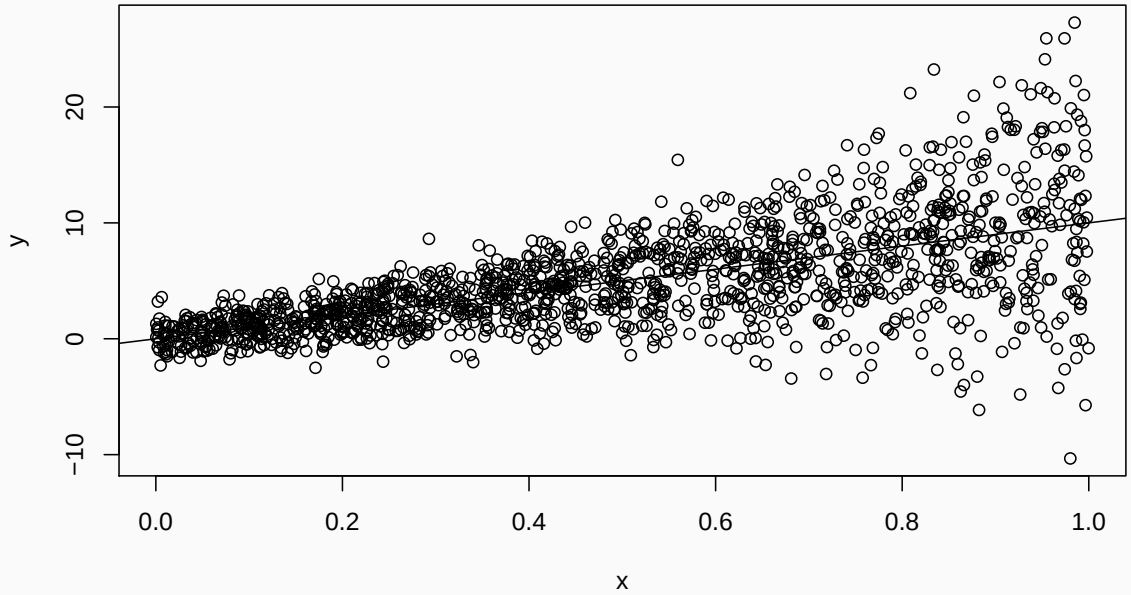
But LS always assumes the variance of that variable is the same:

> 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

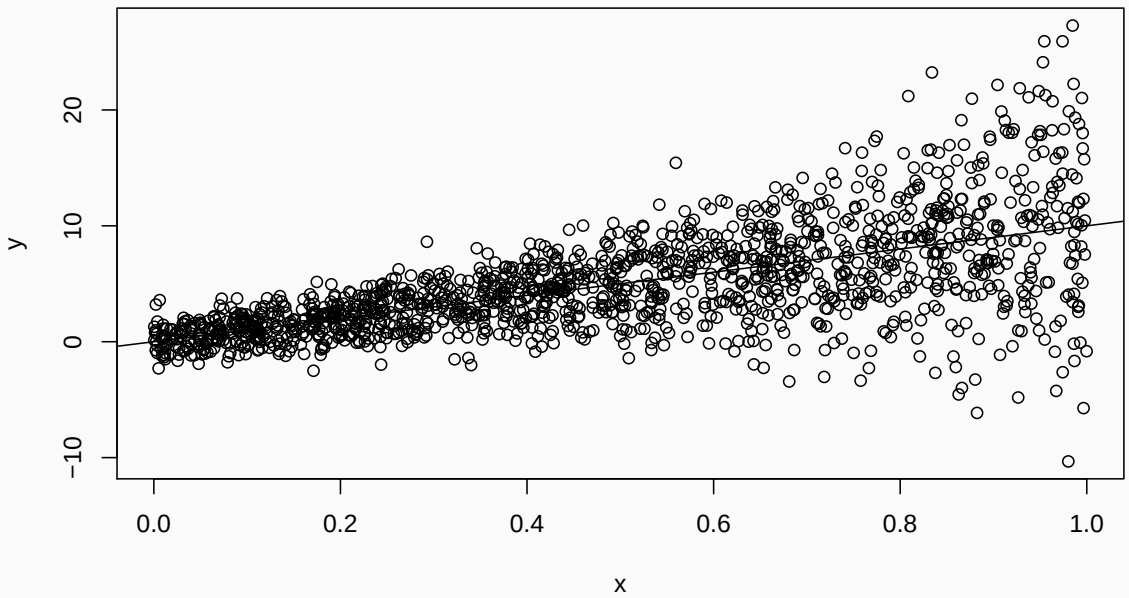
Heteroskedasticity: “Different variance”



A common pattern of heteroskedasticity:
Variance and mean increase together

Here, they are both correlated with the covariate X

Heteroskedasticity: “Different variance”

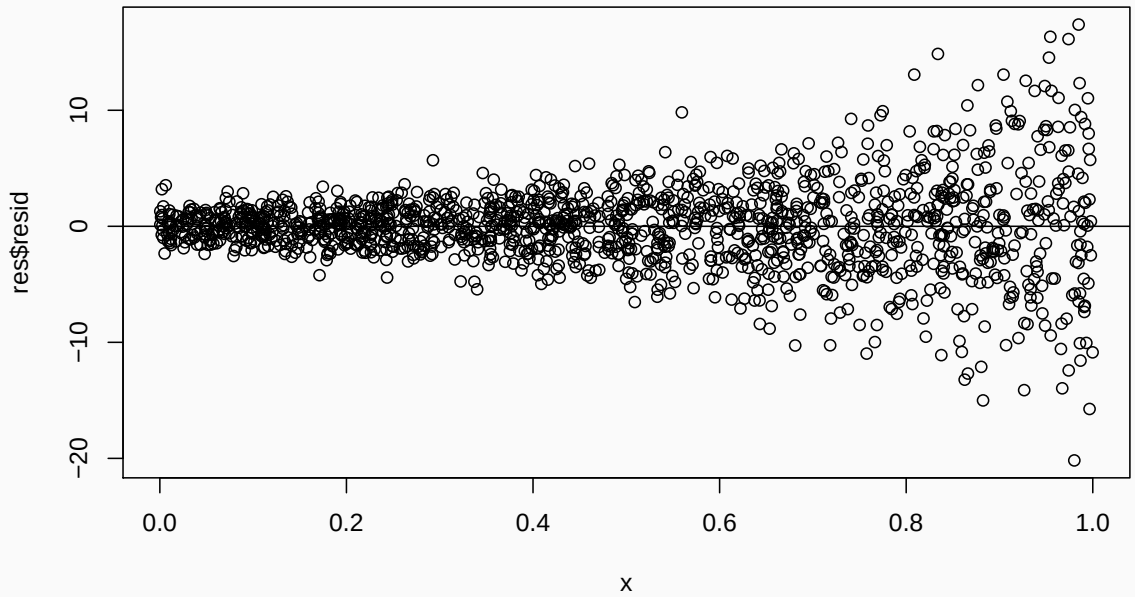


In a fuzzy sense, X is a necessary but not sufficient condition for Y

This is usually an important point substantively.

Heteroskedasticity is interesting, not just a nuisance

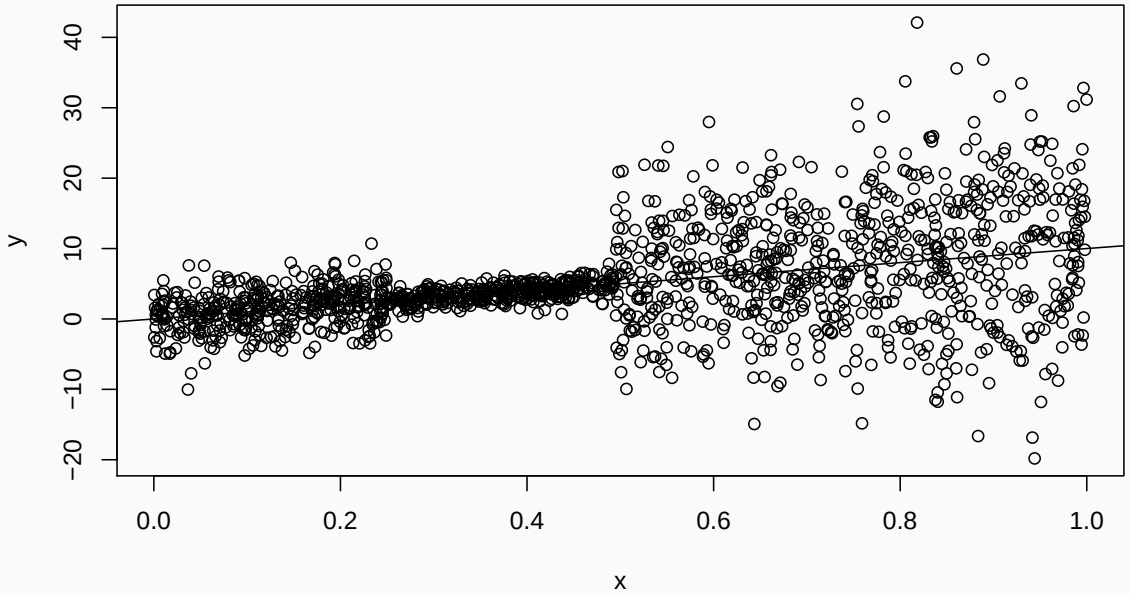
Heteroskedasticity: “Different variance”



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

Look for a pattern. Often a megaphone

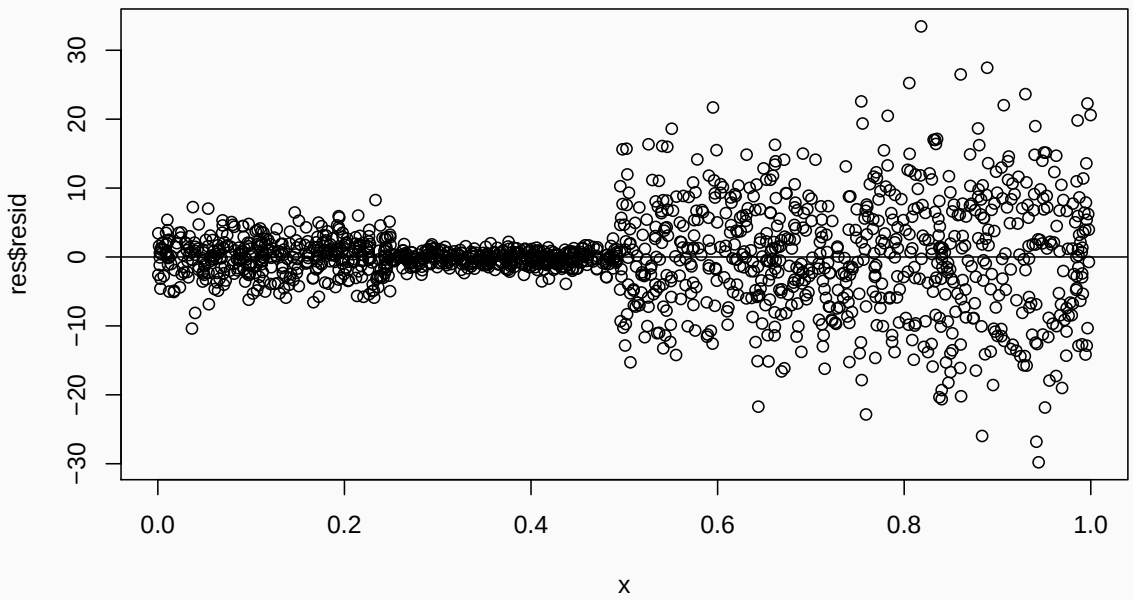
Heteroskedasticity: “Different variance”



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

The Multivariate Normal distribution

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

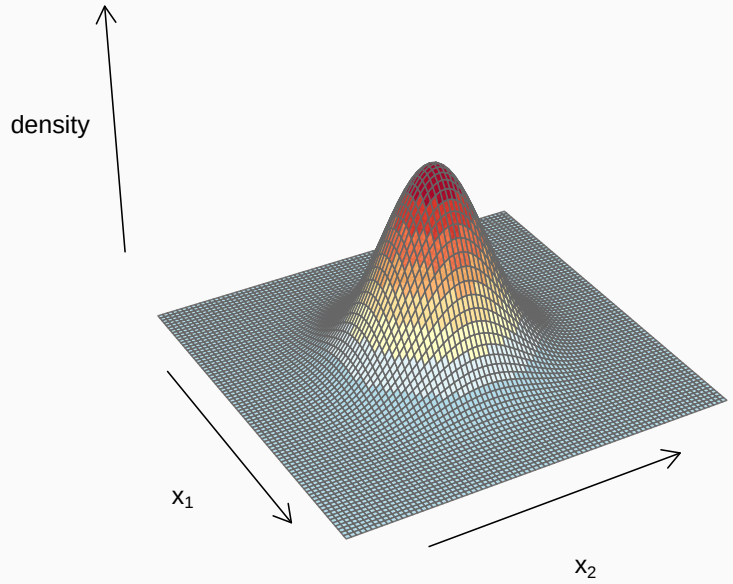
As usual, let μ indicate a mean, and σ a variance or covariance

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

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

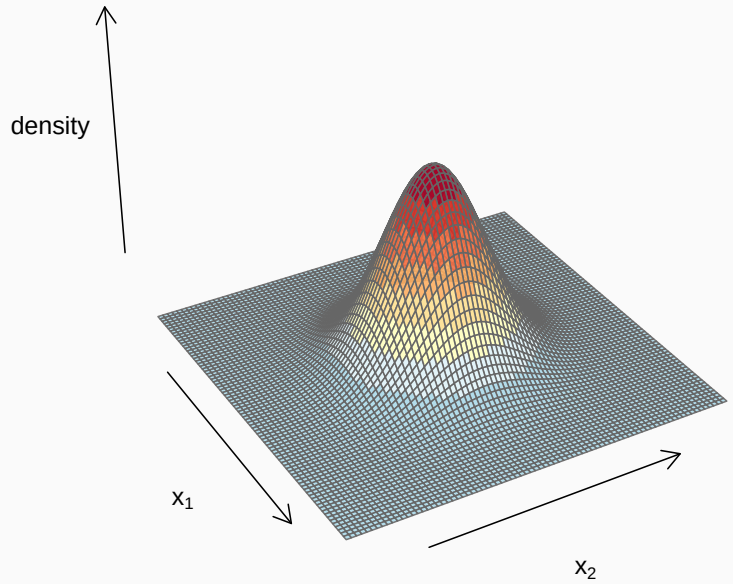
The MVN is more than the sum of its parts:
There is a mean and variance for each variable, and covariance between each pair

The Multivariate Normal distribution



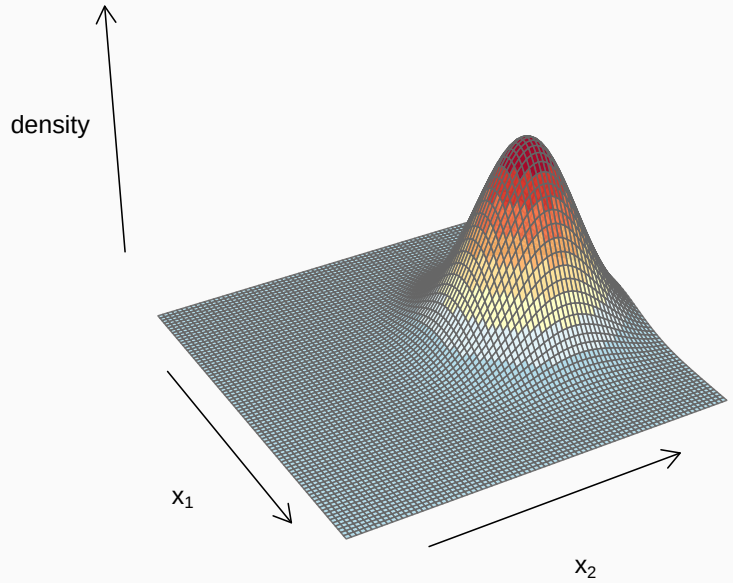
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



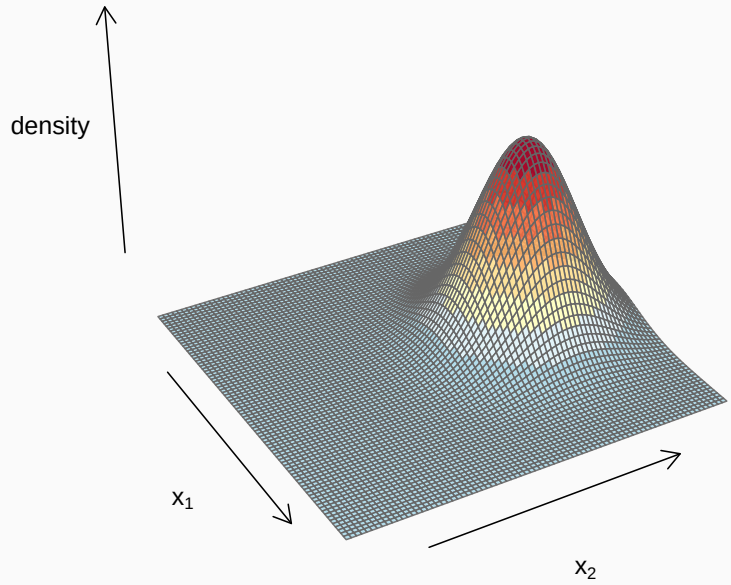
The standard MVN, with zero means, unit variances, and no covariance, looks like a higher dimension version of the normal: a symmetric mountain of probability

The Multivariate Normal distribution



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

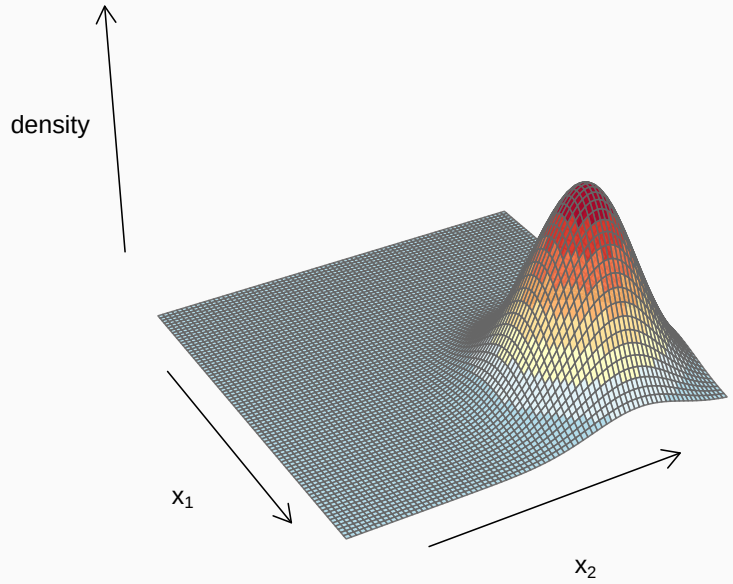
The Multivariate Normal distribution



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

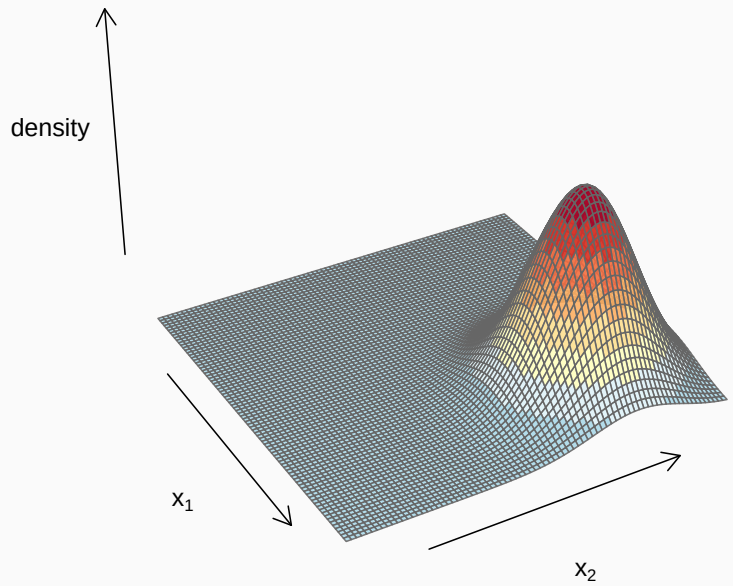
Mean shifts affect only one dimension at a time

The Multivariate Normal distribution



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

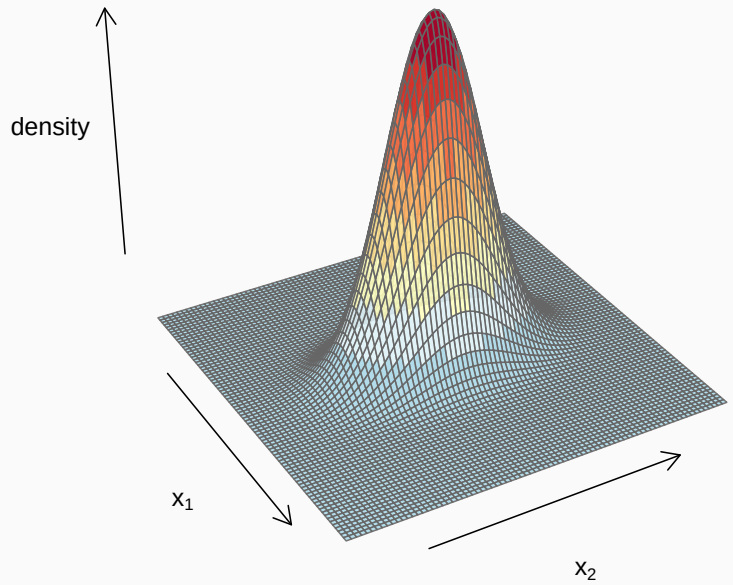
The Multivariate Normal distribution



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

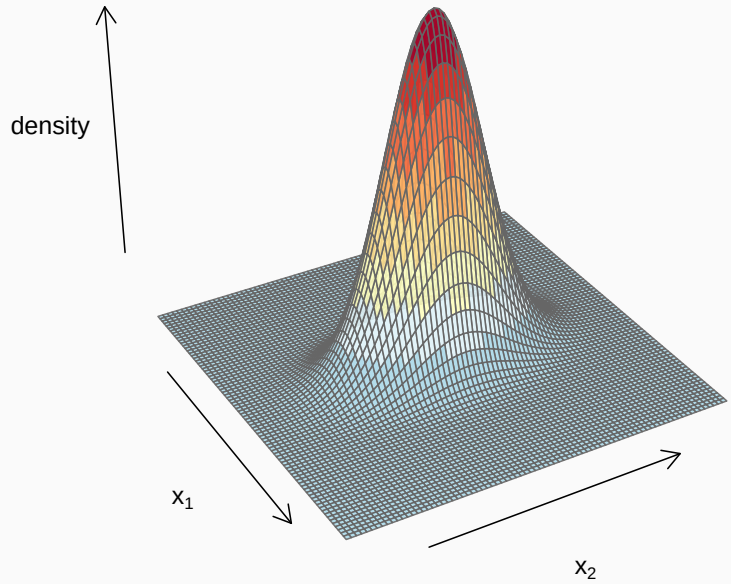
This MVN says the most likely outcome is both x_1 and x_2 will be near 2.0

The Multivariate Normal distribution



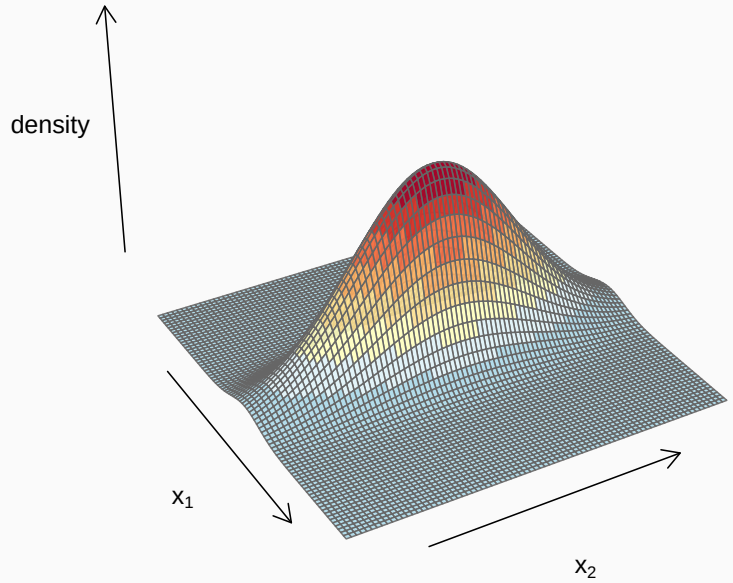
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

The Multivariate Normal distribution



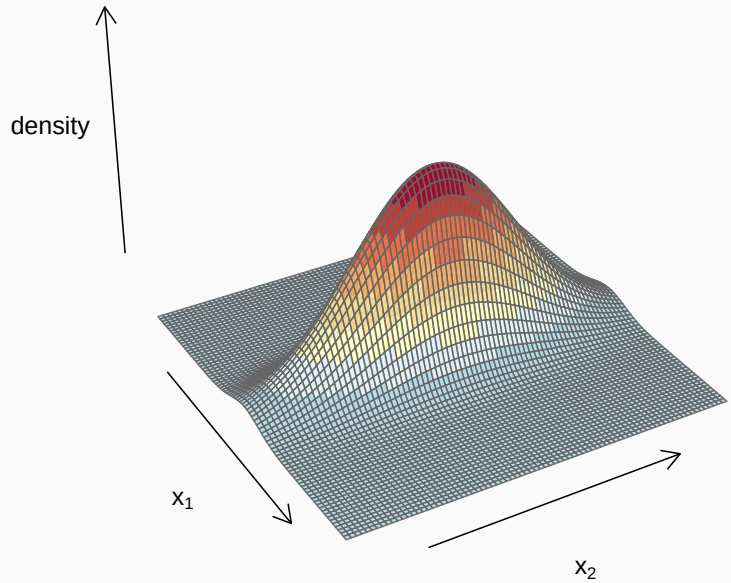
Shrinking the variance of $\mathbf{x}_1$ moves the mass of probability towards the mean of $\mathbf{x}_1$, but leaves the distribution around x_2 untouched

The Multivariate Normal distribution



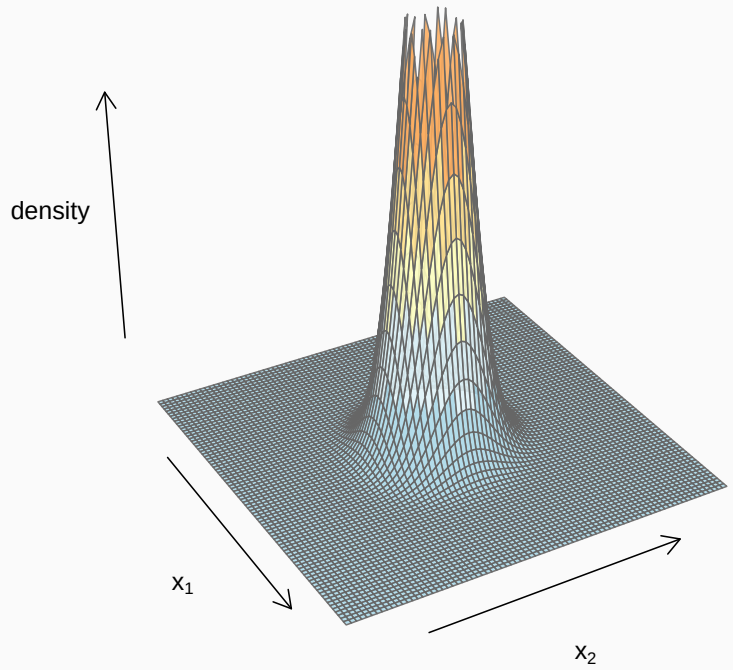
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 3 \end{bmatrix}\right)$$

The Multivariate Normal distribution



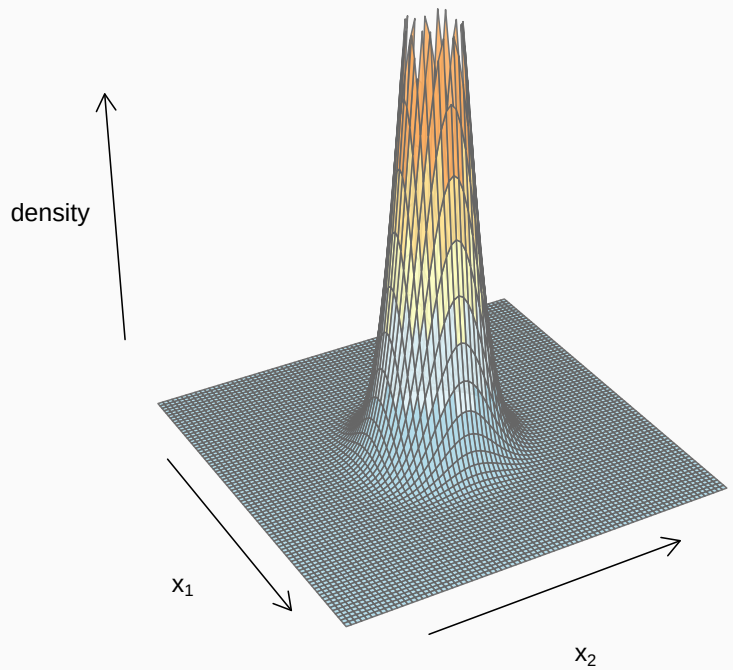
Increasing the variance of x_2 spreads the probability out,
so we are less certain of x_2, but just as certain of x_1 as before

The Multivariate Normal distribution



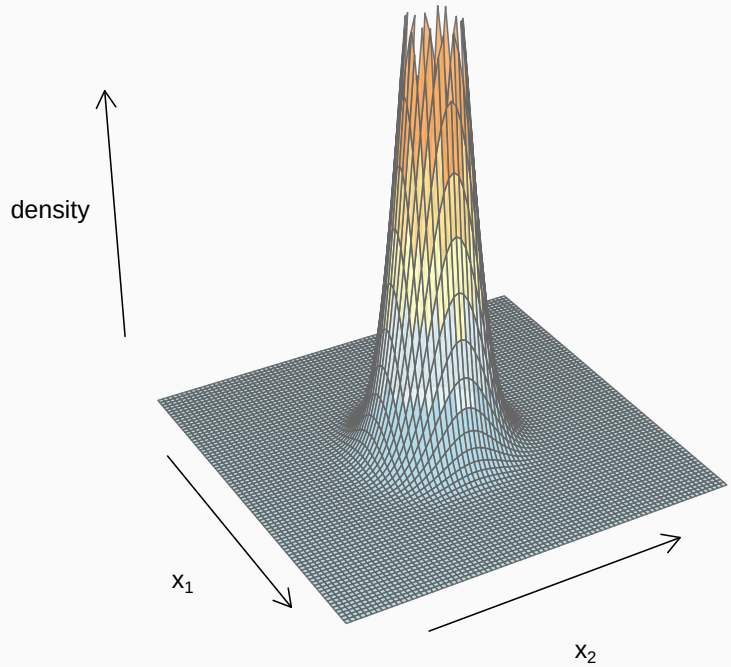
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.33 & 0 \\ 0 & 0.33 \end{bmatrix}\right)$$

The Multivariate Normal distribution



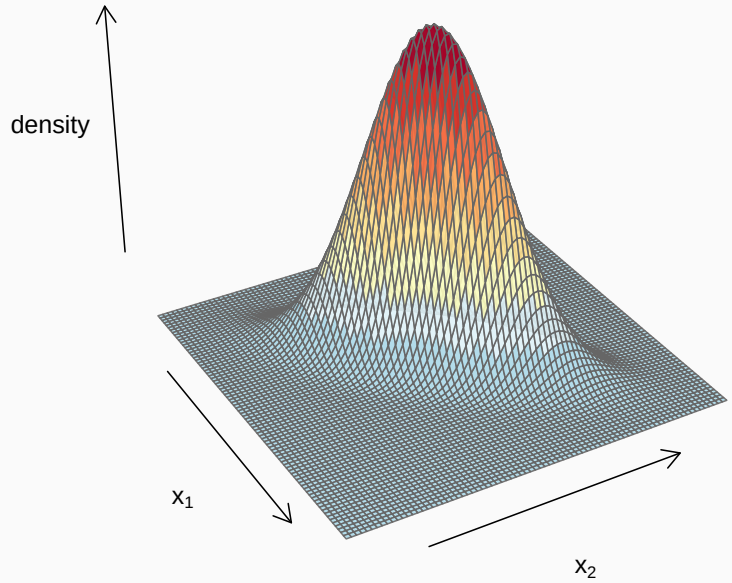
If the variance is small on all dimensions, the distribution collapses to a spike over the means of all variables

The Multivariate Normal distribution



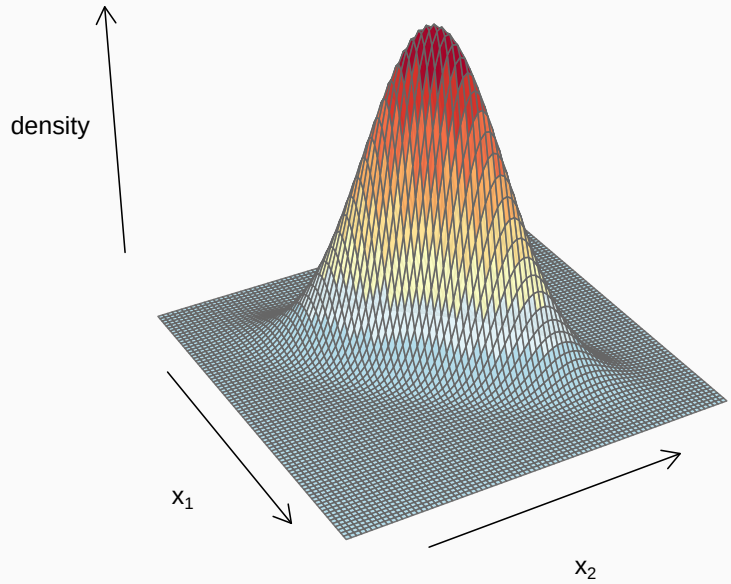
In this case, we are fairly certain of where all our variables tend to lie

The Multivariate Normal distribution



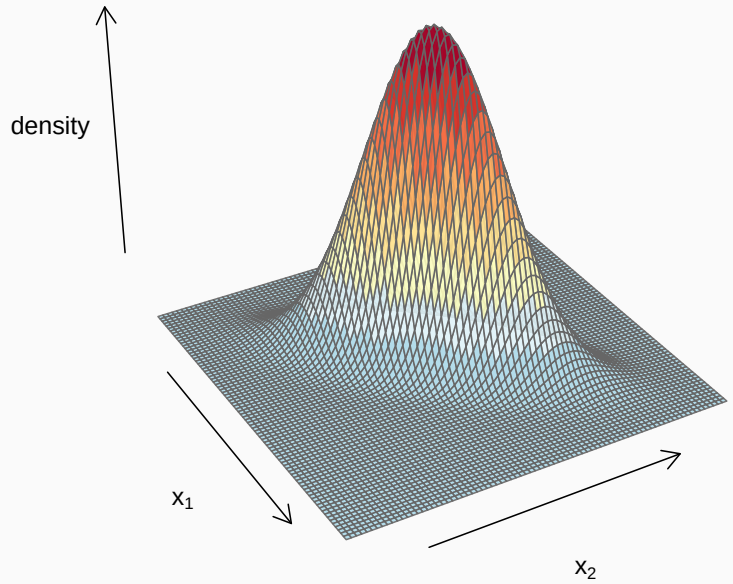
$$\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathcal{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix} \right)$$

The Multivariate Normal distribution



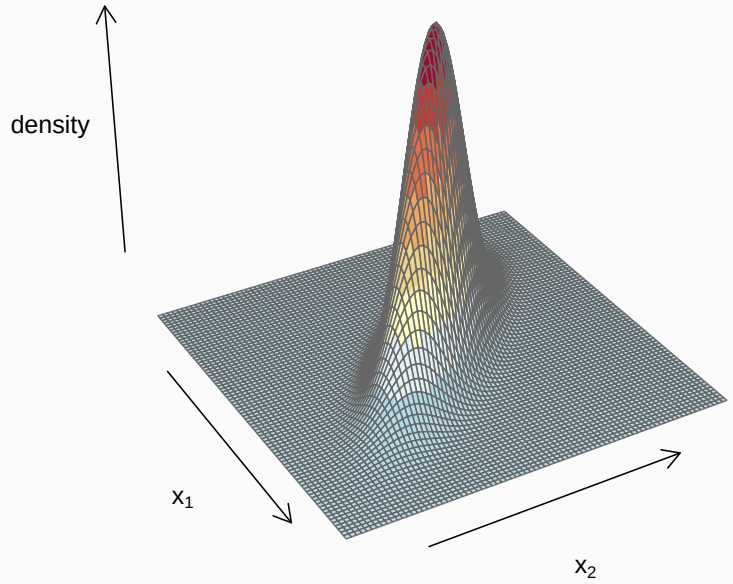
In this special case, with unit variances, the covariance is also the correlation, so our distribution say x_1 and x_2 are correlated at $r = 0.8$

The Multivariate Normal distribution



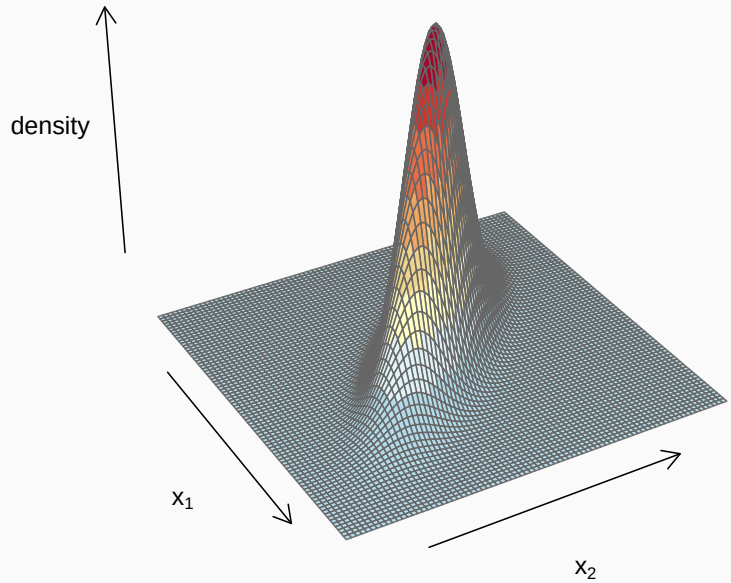
A positive correlation between our variables makes the MVN asymmetric,
with greater mass on likely combinations

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

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

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

Unpacking σ^2: homoskedastic case

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

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

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	$\text{rank}(\mathbf{X}) = k, k < n$	

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	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$	

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	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$

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	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathbb{E}(\boldsymbol{\varepsilon}) = 0$	Biased, even as $N \to \infty$
4	No serial correlation	$\mathbb{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	Unbiased but ineff. se's wrong
5	Homoskedastic errors	$\mathbb{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$	

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	$\mathrm{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	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	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$	

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

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

So long as

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

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

$\rightarrow \beta_{\mathrm{LS}}$ is unbiased and a linear function of 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)