Essex Summer School in Social Science Data Analysis
Panel Data Analysis for Comparative Research

Course Introduction / Review of Linear Regression and Simulation

Christopher Adolph

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

What are Panel Data?

Economic performance of N countries over T quarters

Crime rates in N regions over T years

Opinion of N persons surveyed across T periods

Vote share of governing coalition in N countries over T elections

These are examples of panel or time series cross-section (TSCS) data

In this class, panel and TSCS used interchangeably
for repeated observations on a set of units

Sometimes, TSCS refers to sequential observation on different units,
or to a repeated panel with small N and large T

Likewise, in many fields, panel data are often large N & low T

Will consider this case in detail towards the end of the course, but our initial focus is modeling of continuous outcomes in panels with at least medium T

Panel Data with Medium T (e.g., Political Science)

Excepting survey data, political science datasets often have
fixed and small N, "medium" but expandable T, and continuous outcomes

This combination appears in other social sciences as well,
and forces us to think hard about three issues:

1. Temporal dependence: The past plays a strong role in politics & society. Overlooking or misspecifying dynamics leads to serious errors in inference

2. Unit heterogeneity: Many unobserved differences among our units. Serious danger of omitted variable bias. Proper use of panel tools can mitigate this danger.

3. Heteroskedasticity: In linear models, the variance of disturbances may itself vary, either across units or over time

Can borrow many techniques from time series econometrics, but often our focus (and problems) differ. Economists usually have limited T, large N, and lower(sometimes) temporal dependence.

Goals of the course

By the end of the course, you should be able to:

- Select the appropriate time series or panel model for your data

- Demonstrate the appropriateness of your model
using various tests for heteroskedasticity and serial correlation

- Estimate your model and interpret the results for a broad audience
in terms of the quantities of substantive interest

- Show readers the time series or panel structure of your data,
and how a proper understanding of that structure affects model interpretation

Structure of the course

Part 1: Review of Fundamentals

Basic tools for estimating and understanding models in the course; 1.5 days

Part 2: Modeling Time Series Dynamics

Stochastic Processes / Lagged DV / ARIMA / Cointegration; 3.5 days

Part 3: Modeling Panel Heterogeneity

Random effects / Fixed effects / Panel GMM / Heteroskedasticity; 3.5 days

Part 4: Advanced Topics (examples)

In-Sample Simulation / Multiple Imputation / Binary TSCS; 1.5 days

Motivating Examples

What will you get out of the course?

Three motivating examples from my own work. . .

1. Comparative Inflation Performance

2. African Labor Standards

3. Homicide Among Dating Partners in the US

See separate slide shows (linked to course page)

Outline for Topic 1

1. Review of linear regression (notation, estimation)

2. Desired properties of estimators (bias, efficiency, consistency)

3. Assumptions underlying linear regression (Gauss-Markov Theorem)

4. Heteroskedasticity & simple "fixes" (robust standard errors)

5. Estimation and model selection with MLE

6. Clear presention of substantive results through simulation

More topics above than we can cover comprehensively;
read ahead and ask questions so we can focus on areas of most interest/need

Review of simple linear regression

Recall the linear regression model in scalar form

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

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

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

Note that:

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

The systematic component is: $\mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}$

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

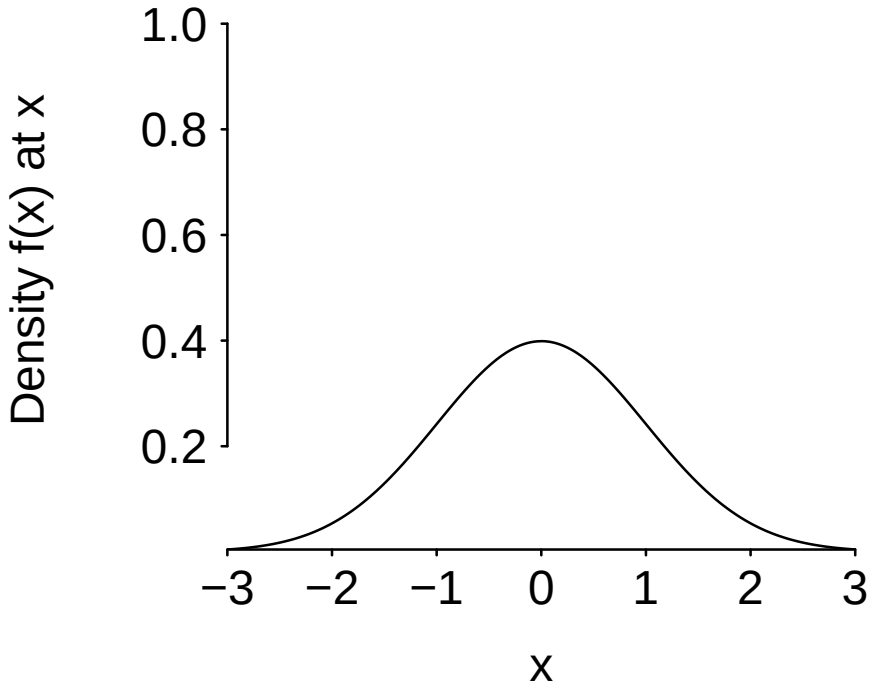
Aside: The Normal (Gaussian) distribution

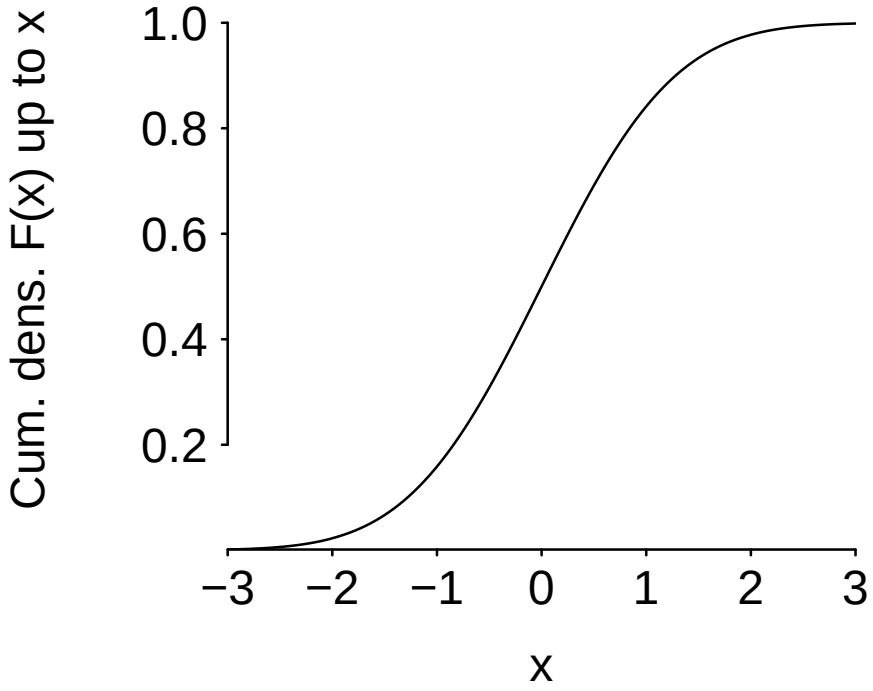
$$f_{\mathcal{N}}(y_i|\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[\frac{-(y_i-\mu)^2}{2\sigma^2}\right]$$

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

Many researchers implicitly or explicitly assume their data are Normally distributed

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





Review of simple linear regression

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

$$\begin{aligned}
\sigma^2 &= \mathbb{E}\left(\left(\varepsilon_i - \mathbb{E}(\varepsilon_i)\right)^2\right) \\
&= \mathbb{E}\left(\left(\varepsilon_i - 0\right)^2\right) \\
&= \mathbb{E}(\varepsilon_i^2) = \sum_{i=1}^{n} \varepsilon_i^2/n \\
\sigma &= \sqrt{\sum_{i=1}^{n} \varepsilon_i^2/n}
\end{aligned}$$

Note that this is the square root of the mean of the squared errors (RMSE)

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

σ is how much we expect y_i to differ from its expected value, $\beta_0 + \sum_k \beta_k x_{ki}$, on average

Estimates of σ are the basis of many measures of goodness of fit

Linear Regression in Matrix Form

Scalar representation:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

Equivalent matrix representation:

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

Which uses these matrices:

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

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.

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

Linear Regression in Matrix Form

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

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

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

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

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

$'$ (or $^{\mathrm{T}}$) is the transpose operator: it flips a matrix along the main diagonal

Linear Regression in Matrix Form

Recall $\mathbb{E}(\varepsilon_i \varepsilon_j) = 0$ 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
to relax the assumptions

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

$$\arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \varepsilon_i^2 = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$$

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

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

Linear Regression in Matrix Form

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

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

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

Linear Regression in Matrix Form

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

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

Just like a bunch of scalar derivatives stacked together:

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

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

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

Linear Regression in Matrix Form

A similar pattern holds for quadratic expresssions.

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

And the vector analog 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 than,
and at times at variance with, the plain English meaning.

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

Although it seems obvious that we always prefer an unbiased estimator
if one is available, a little thought shows this is not the case [diagram on board]

Why? 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 could be utterly useless.
One unbiased estimate of the time of day: a random draw from the numbers 0–24.

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 the bias and variance of estimator.

A biased estimate with low variance may be "better" (more useful)
than an unbiased estimate with high variance

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:

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

If $\mathbb{E}(\hat{\beta}) = \beta$ (no bias), MSE reduces to $\text{var}(\hat{\beta})$

$\sqrt{\text{MSE}}$ (or RMSE) 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:
If an estimator converges to the truth as the number of observations grows it is *consistent*

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

Of great concern to many econometricians and arguably important as N gets big
e.g., as in survey data, where N is typically a matter of research design decisions

Not as great a concern in, say, comparative politics (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 much less with consistency

. . . that said, application of some panel data methods to some panel data structures will produce inconsistent estimates, as we shall see

What can go wrong in a linear regression?

Even if your data are sampled without bias from the population of interest, and your model is correctly specified (contains the "right" control variables), several data and design problems can violate the linear regression assumptions

In order of declining severity, these are:

Perfect collinearity

Endogeneity of covariates

Heteroskedasticity

Serial correlation

Non-normal disturbances

Perfect Collinearity

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

Matrix inversion – and thus LS regression – is impossible here

Your stat package will return an error and/or drop covariates randomly
(don't report such a model – respecify to remove collinearity first)

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 in applied research – except in panel data, as we will see!

"Partial" Collinearity

What if our covariates are correlated but not perfectly so?

Then they are *not* linearly dependent

The regression coefficients are identified in a linear algebra sense:
a unique estimate exists for each β_k

Regression with partial collinearity is unbiased & efficient, all else equal

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 standard errors are *correct*

They are not a statistical problem to be fixed;
instead, the data are too limited for the model
or the model too ambitious for the sample

"Partial" Collinearity

"Partial" Collinearity is actually an oxymoron

Social scientists sometimes inappropriately call this "multicollinearity"

In mathematics, 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 standard errors?
Then you lack sufficient data to precisely answer your research question

$\rightarrow$ expand the data or narrow the 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

$\mathbf{y}$ 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 reverse or reciprocal causation?

Exogenous & endogenous variables

That is, what if there is reverse or reciprocal causation?

Very common in political science and other social sciences:

- campaign spending and share of the popular vote
- arms races and war, etc
- students' measured aptitude and effort
- individual health and income

In these cases, $\mathbf{y}$ and $\mathbf{X}$ 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 $N \to \infty$

Exogenous & endogenous variables

How do you identify a causal relationship when $\mathbf{y}$ and $\mathbf{X}$ are both endogenous?

$\rightarrow$ Revise your research design:

1. Conduct a lab and/or field experiment with random assignment of treatment

2. Find a natural experiment ("as-if" random assignment)

3. Find an exogenous "version" of $\mathbf{x}$

4. Find an instrument for $\mathbf{x}$: a variable $\mathbf{z}$ that affects $\mathbf{y}$ only through $\mathbf{x}$

5. Find a deterministic "discontinuity" in real-world application of a treatment and apply a regression discontinuity design

Sometimes panel data can help:
especially if we can sequence cause and effect in time (Options 3 & 4),

or compare before and after a naturally occuring intervention
as in difference-in-differences designs (Option 2)

Heteroskedasticity: "Different variance"

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

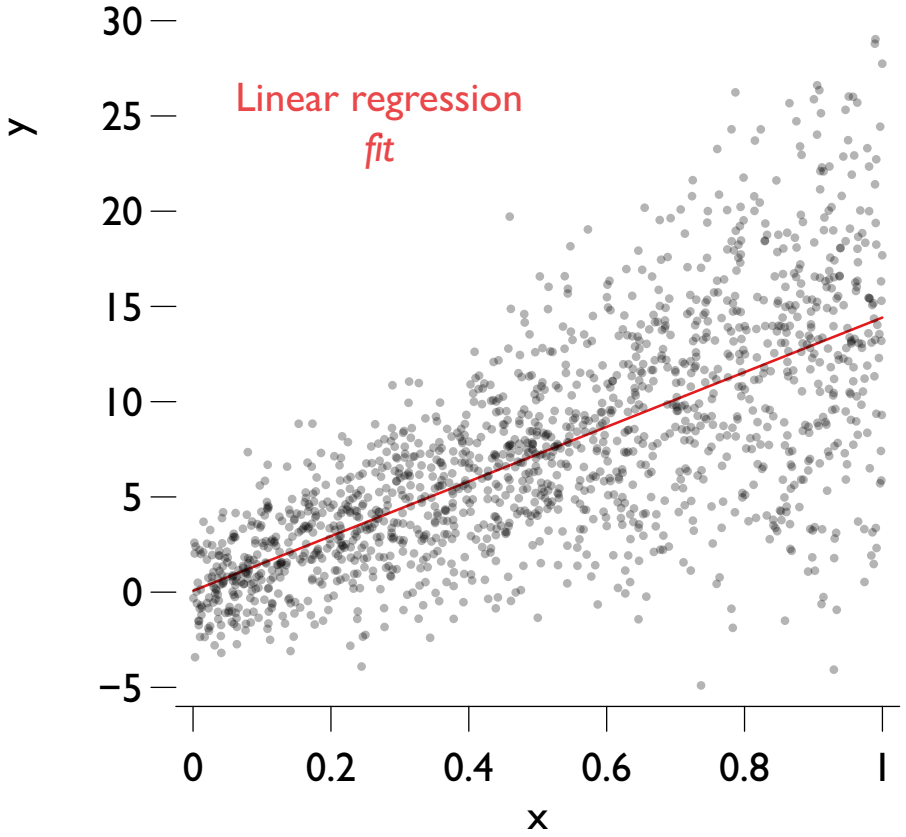
$\mathbf{y}$ could be any linear function of $\boldsymbol{\beta}$ and $\mathbf{X}$

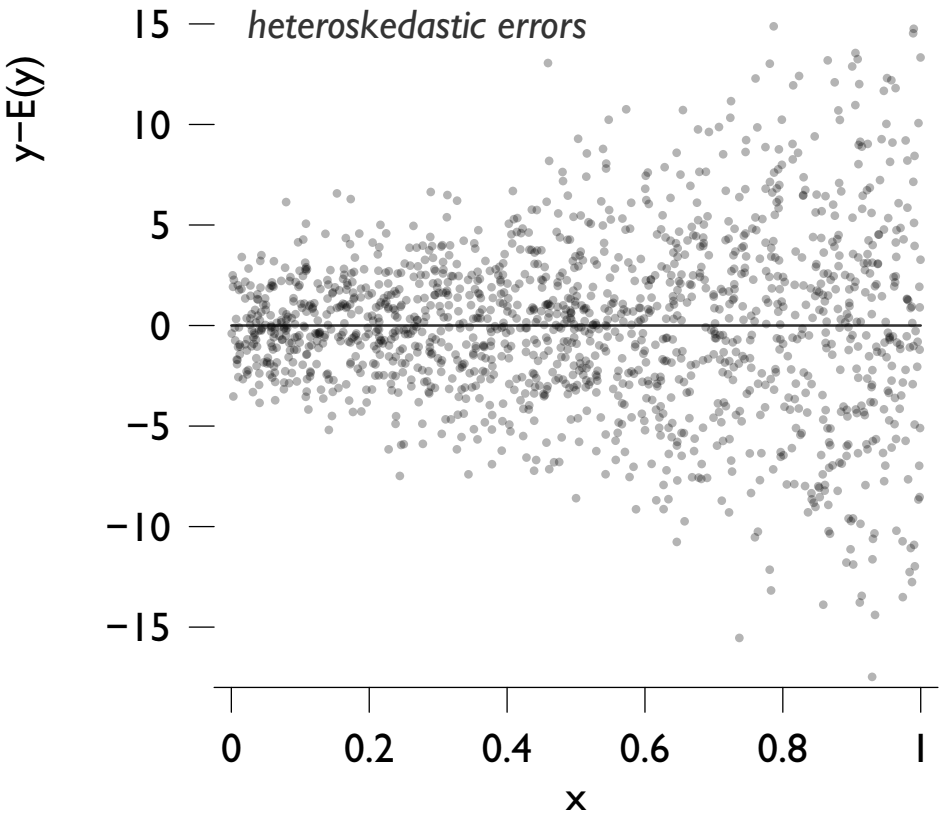
But LS always assumes the variance of that variable is the same:

σ^2, a constant

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

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



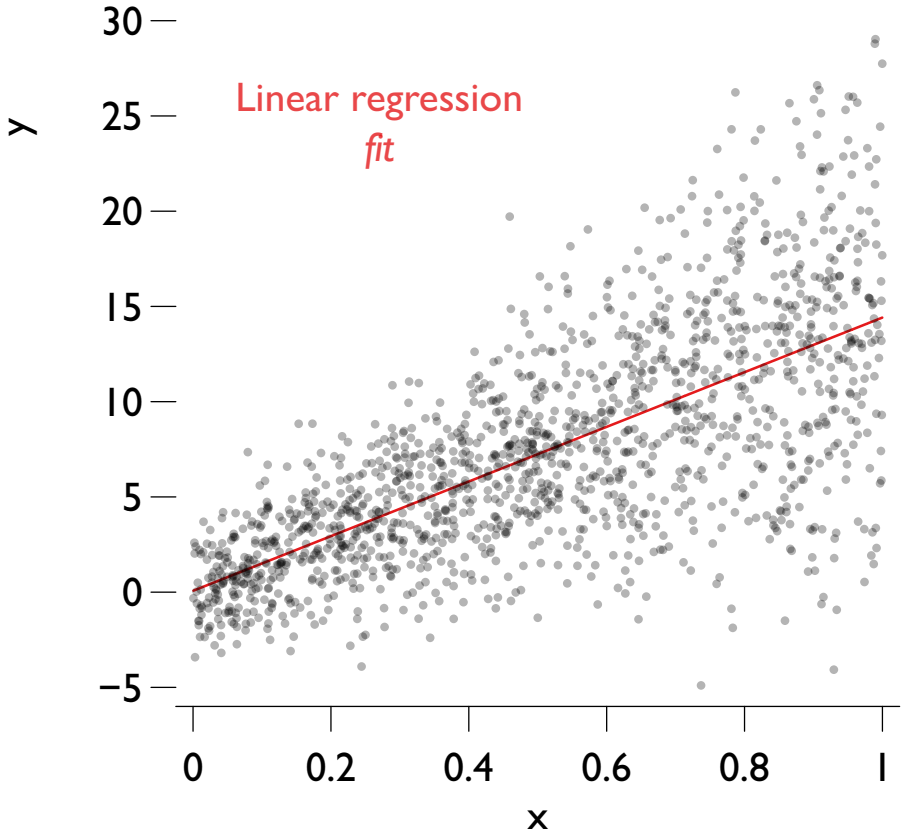


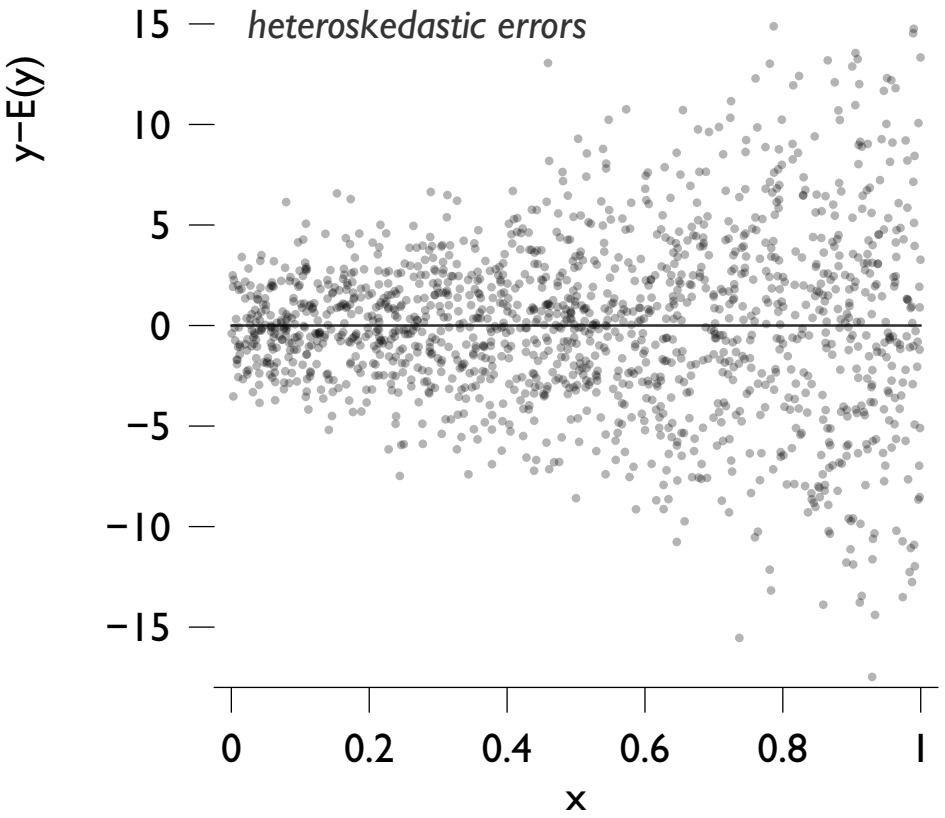
A common pattern of heteroskedasticity: variance and mean increase together

Here, they are both correlated with the covariate x

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

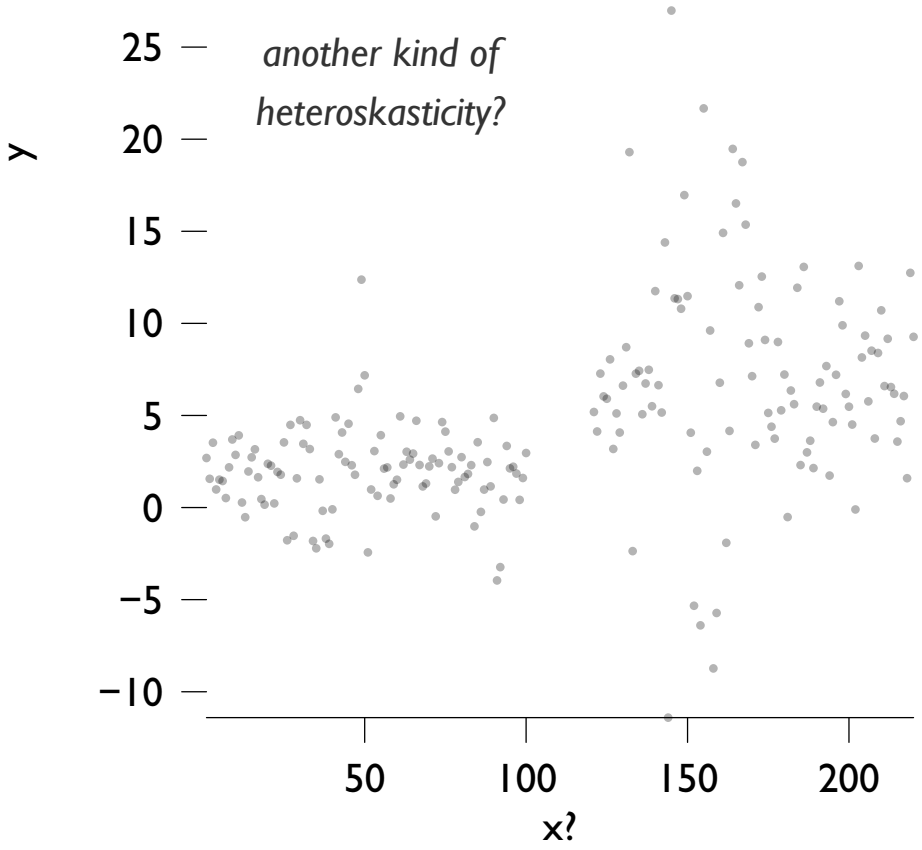
Heteroskedasticity is often substantively interesting, but mistaken for mere nuisance

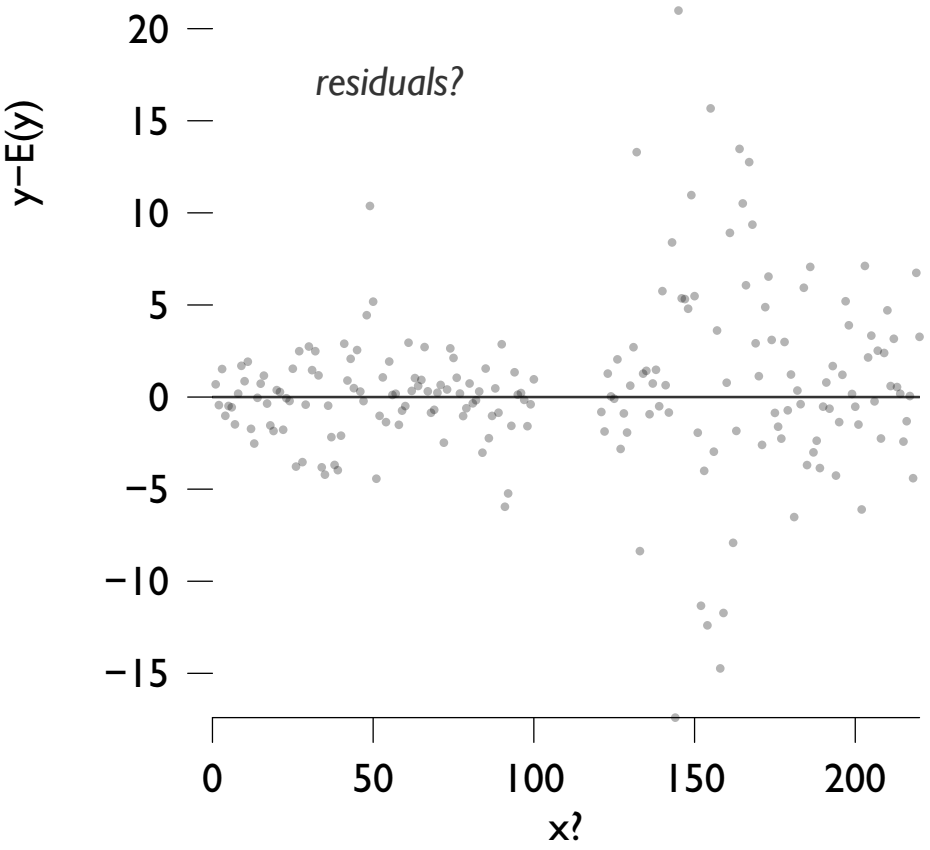




Diagnose heteroskedasticity by plotting the residuals against each covariate:
Look for a pattern, often a megaphone

But other patterns are possible

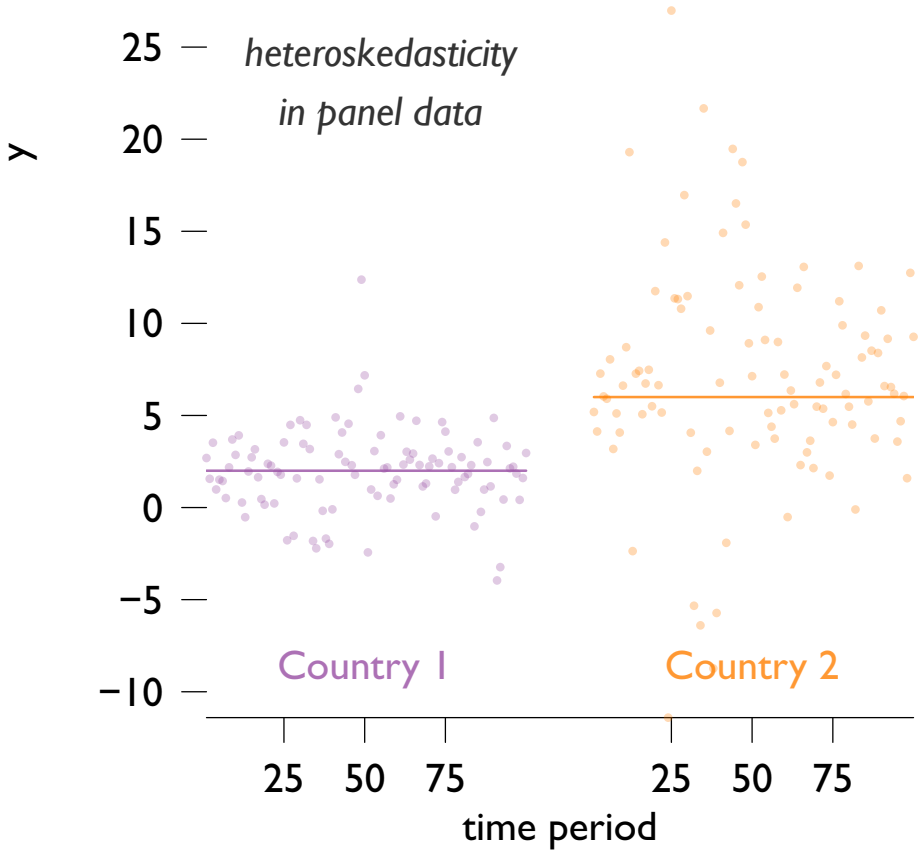


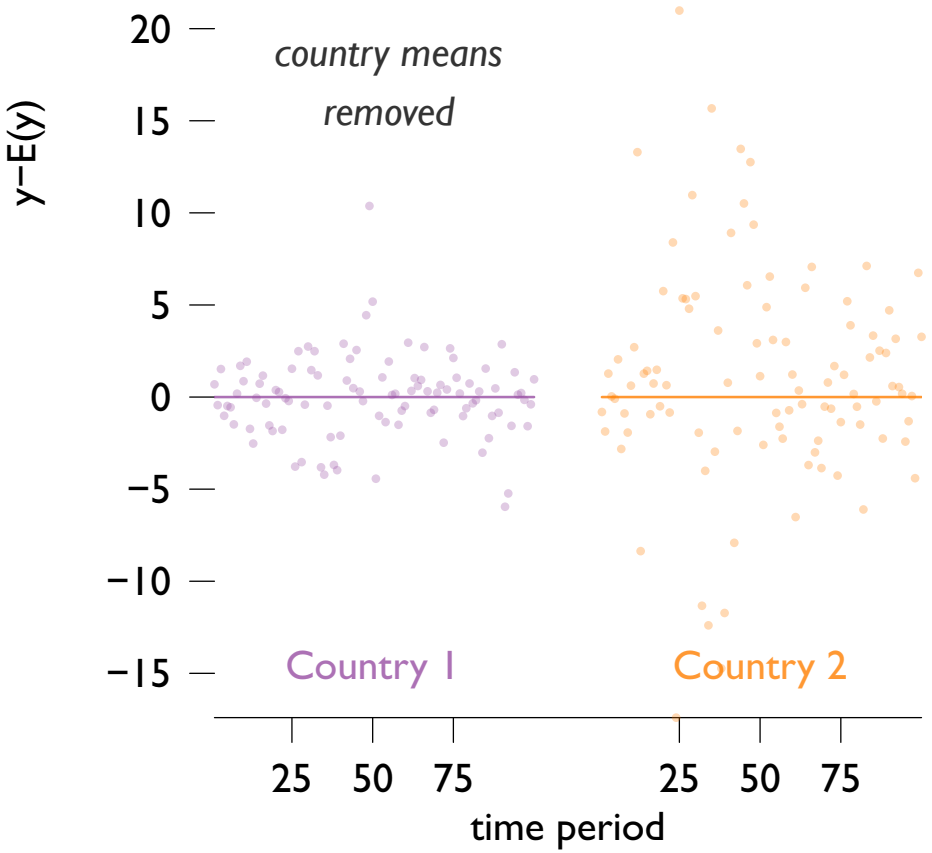


Diagnose heteroskedasticity by plotting the residuals against each covariate:
Look for a pattern, often a megaphone

But other patterns are possible

What do you think is happening in this example?



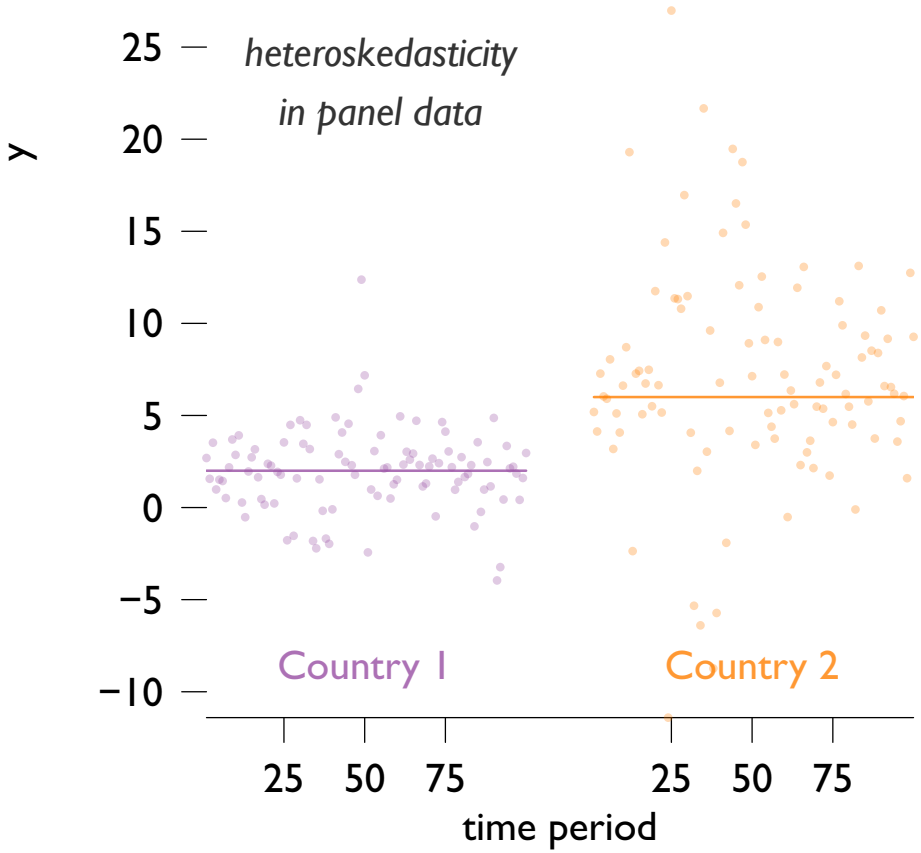


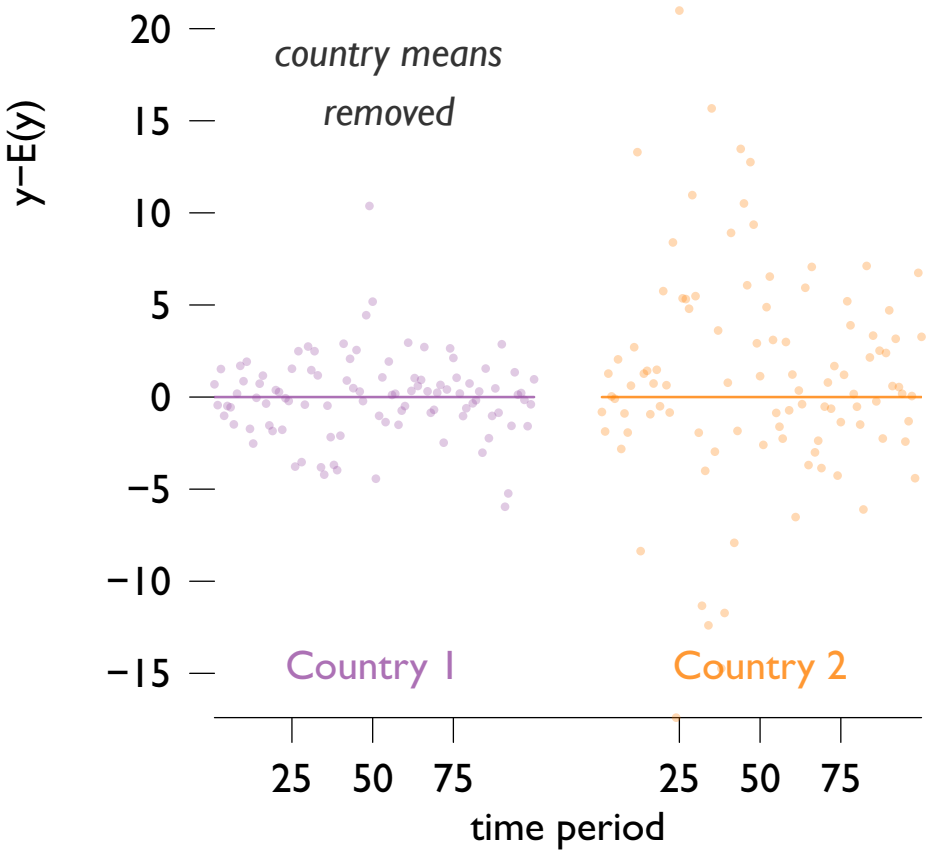
Heteroskedasticity can be more complex in panel datasets, including:

(1) variances differ by cross-sectional unit: *country 2 has higher variance*

(2) variances differ by time periods: *periods 25–50 have higher variance*

(3) lingering heteroskedasticity after large random errors: *present but subtle*





More on panel heteroskedasticity later in the course. . .

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 $\mathbf{x}_2$

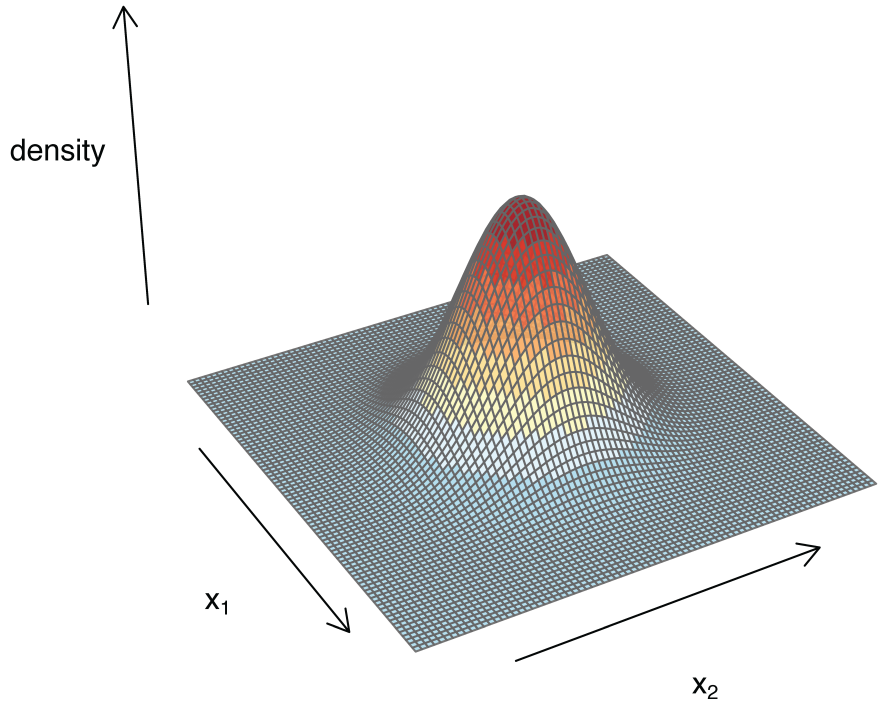
As usual, let μ indicate a mean, and σ a variance or covariance

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

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

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

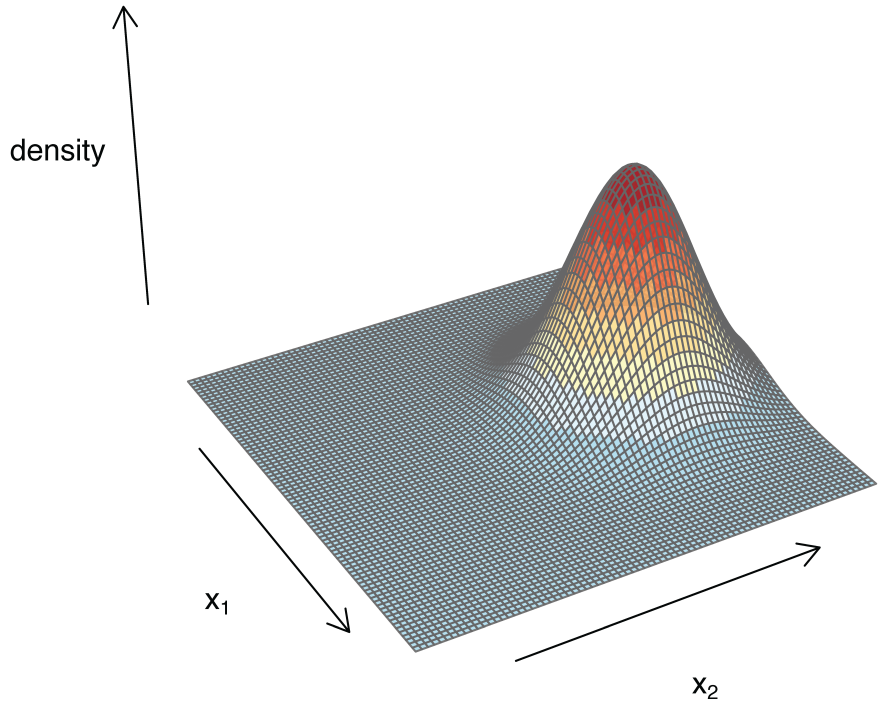
The Multivariate Normal distribution



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

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

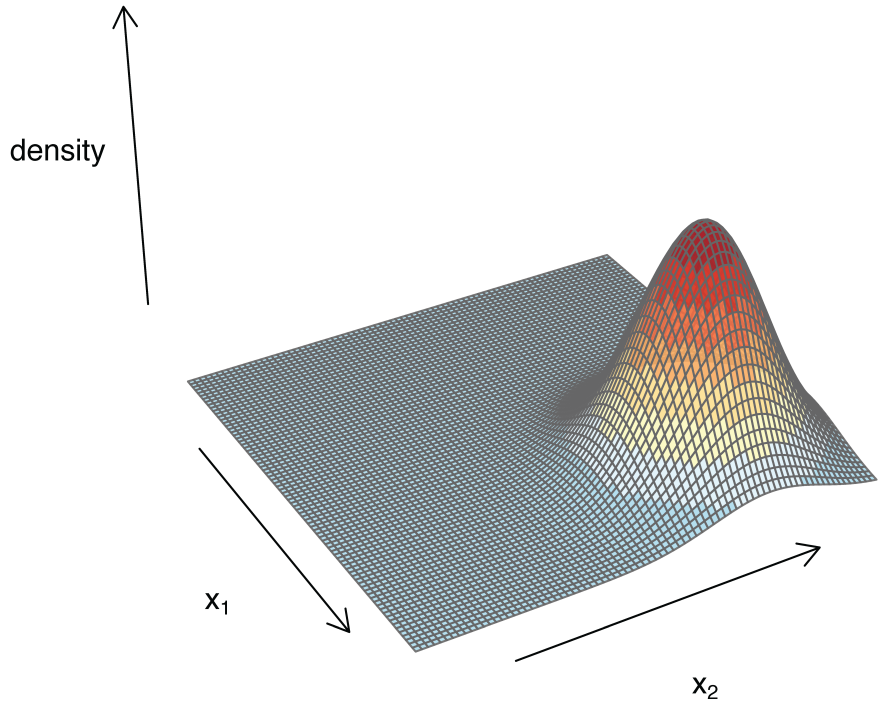
The Multivariate Normal distribution



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

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

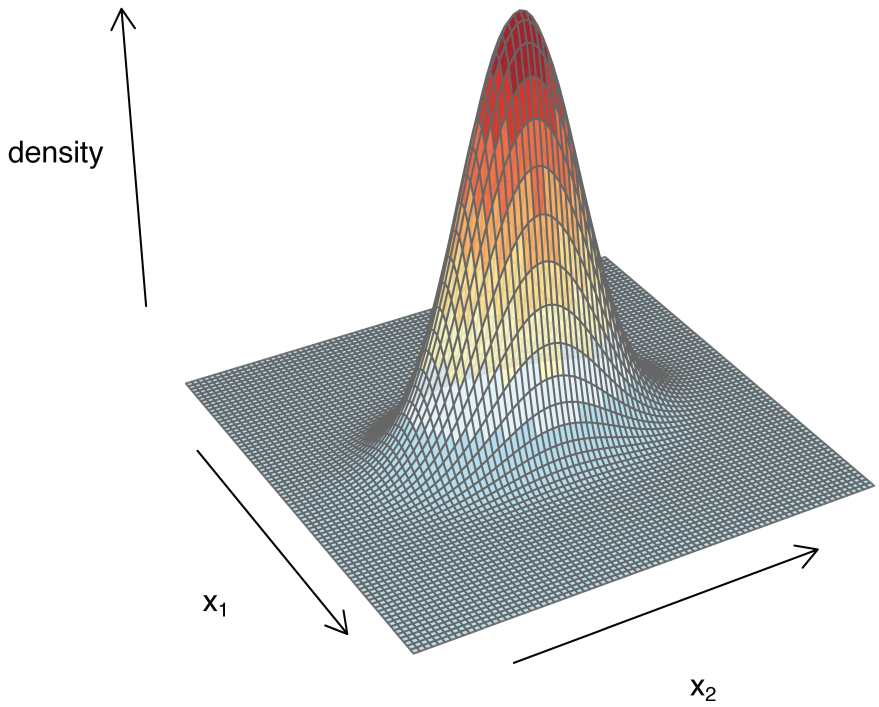
The Multivariate Normal distribution



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

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

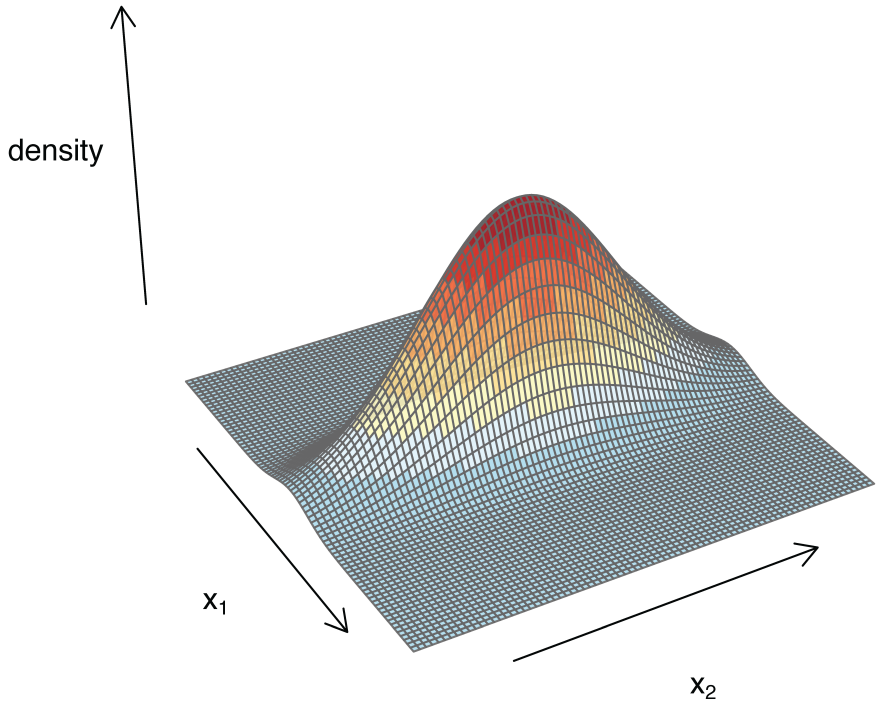
The Multivariate Normal distribution



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

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

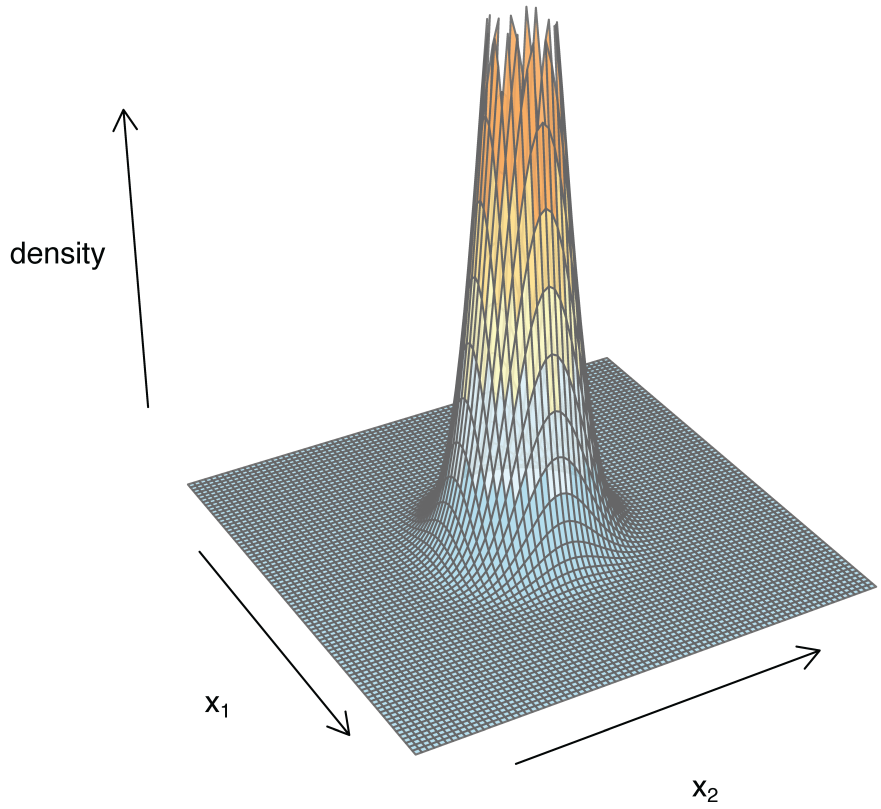
The Multivariate Normal distribution



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

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

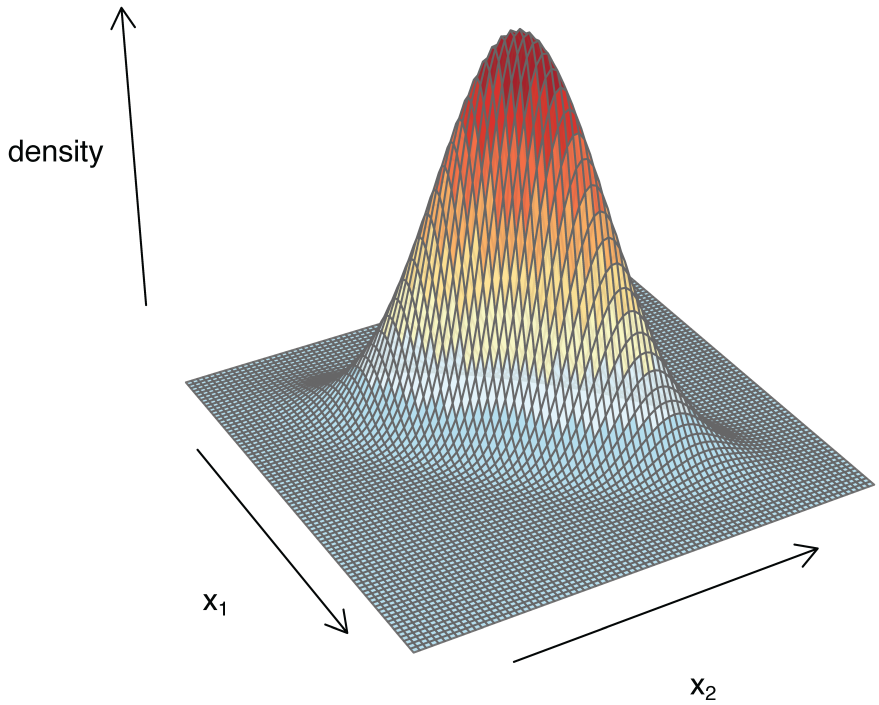
The Multivariate Normal distribution



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

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

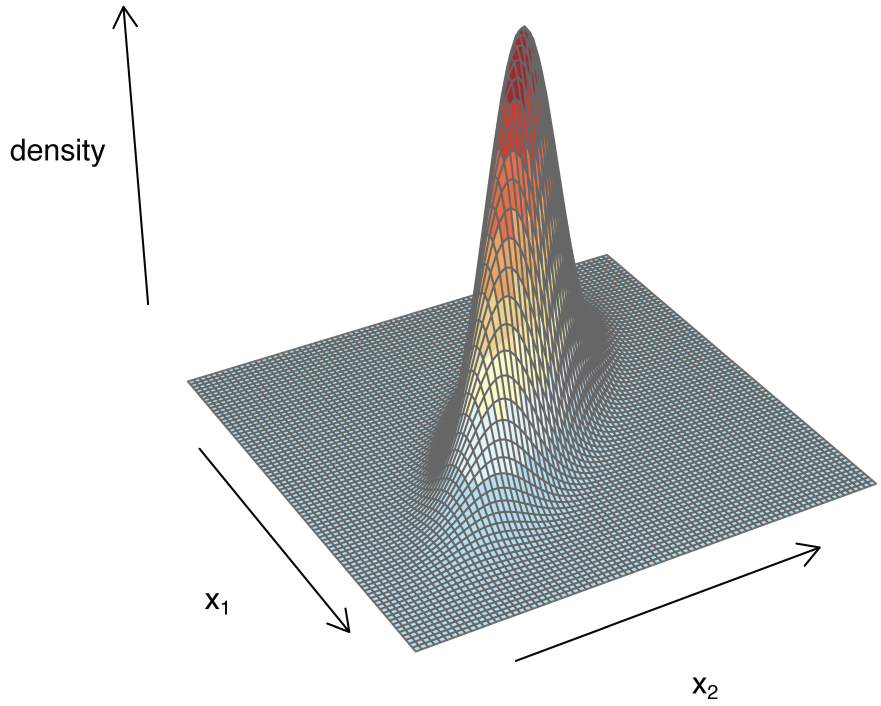
The Multivariate Normal distribution



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

With unit variances, covariance = correlation, so $\mathrm{cor}(\mathrm{x}_1, \mathrm{x}_2) = 0.8$
A positive correlation between our variables makes the MVN asymmetric

The Multivariate Normal distribution



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

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 $\mathbf{x}_i\boldsymbol{\beta}$ plus
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

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

There is 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

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 estimates of standard errors 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 estimates of β are unbiased but inefficient in this case

The LS standard errors *will* be biased, however
(provided some weak conditions hold)

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

Gauss-Markov Conditions

So when is least squares unbiased?

When is it efficient?

When are the estimates of standard errors unbiased?

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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	

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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ 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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathbb{E}(\boldsymbol{\varepsilon}) = 0$	

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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathbb{E}(\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ 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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathbb{E}(\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
4	No serial correlation	$\mathbb{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	$\hat{\beta}$ unbiased but ineff. se's biased
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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathbb{E}(\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
4	No serial correlation	$\mathbb{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	$\hat{\beta}$ unbiased but ineff. se's biased
5	Homoskedastic errors	$\mathbb{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$	$\hat{\beta}$ unbiased but ineff. se's biased
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	$\mathbb{E}(\mathbf{X}\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
3	Disturbances have mean 0	$\mathbb{E}(\boldsymbol{\varepsilon}) = 0$	$\hat{\beta}$ biased, even as $N \to \infty$
4	No serial correlation	$\mathbb{E}(\varepsilon_i \varepsilon_j) = 0, i \neq j$	$\hat{\beta}$ unbiased but ineff. se's biased
5	Homoskedastic errors	$\mathbb{E}(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$	$\hat{\beta}$ unbiased but ineff. se's biased
6	Gaussian error distrib	$\varepsilon \sim \mathcal{N}(0, \sigma^2)$	se's biased 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 assumptions 1–3:

If $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$, 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
- $\mathbb{E}(\boldsymbol{\varepsilon}) = 0$ (regardless of the distribution of $\boldsymbol{\varepsilon}$)

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

$\rightarrow \beta_{\mathrm{LS}}$ 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 A1–A5, the disturbances are normally distributed (A6), 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.

Weighted least squares

Heteroskedasticity makes LS inefficient. Why?

Observations with higher variance in errors contain less information

Observations with lower variance tend to be very close to the LS line

(Example: economic performance or budget data for small & large regions)

We'll get better (more efficient) estimates if we give more *weight* to the latter

Needed:

A measure (or predictor) of $\mathrm{Var}(\varepsilon_i)$ for each i

A method for weighting observations in least squares estimation

Weighted least squares

The method is simple: just add weight terms to the estimator

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

where $\mathbf{W}$ is a diagonal matrix with weights w_i on the diagonal

What are the weights? They are (proportional to) the standard error for each y_i

Ideally, the weights are defined such that

$$\begin{aligned} \varepsilon_i &\sim \mathcal{N}(0, \sigma_i^2) \\ w_i^2 &= \frac{1}{\sigma_i^2} \end{aligned}$$

$\rightarrow$ the larger the weight w_i, the smaller the variance of ε_i, the more information in i

We won't often have weights. (Exception: Estimated dependent variables)

Adjusting standard errors for heteroskedasticity

Suppose we diagnose, or suspect, heteroskedasticity, but have no weights

We cannot use WLS, and thus rely on the less efficient LS estimates

But we can try to get standard errors approaching the WLS se's

Recall the standard errors from LS are the square roots of the diagonal elements of

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

So if σ^2 varies by i, these will be badly estimated

A "heteroskedasticity robust" formula for the Var-Cov matrix is:

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}(\sum_i \hat{\varepsilon}_i x_i' x_i)(\mathbf{X}'\mathbf{X})^{-1}$$

Adjusting standard errors for heteroskedasticity

$$\hat{\mathrm{V}}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}(\sum_i \hat{\varepsilon}_i x_i' x_i)(\mathbf{X}'\mathbf{X})^{-1}$$

SE's calculated from this equation are known by many names:

- White standard errors
- robust standard errors
- sandwich standard errors
- heteroskedastcity consistent standard errors

A vast literature exploring refinements to this basic concept

Yet robust standard errors are always a second best approach. Why?

More satisfying solutions to heteroskedasticity

Wait a minute: Is heteroskedasticity just a "problem" to be "fixed"?

Or is it an interesting feature of the data?

- Suppose we were studying the incomes of American households,
- US median real incomes have been stagnant for the past three and a half decades
- Does this mean that the median family is just as well-off?
- Not necessarily: Jacob Hacker notes the variance in incomes is rising over time
- $\uparrow$ in economic risk, even though the expected economic state is unchanged

Many theories predict the amount of "instability" in some outcome,
rather than (or in addition to) the mean of the outcome

These theories could be tested by modeling heteroskedasticity directly

General purpose inference

We've reviewed two estimators so far: LS and WLS

In intro stat classes, you were likely given any estimator you needed

What if you want to use (or even create) a new estimator,
and don't know how to calculate $\hat{\boldsymbol{\beta}}$?

We'd like a general method that always cranks out $\hat{\boldsymbol{\beta}}$, even for a brand new model

E.g., what if we design a model of heteroskedastic data,
or serially correlated data – how do we estimate the model?

We also need to obtain standard errors of our estimates,
which come from the variance-covariance matrix of parameters, $\mathrm{V}\left(\hat{\boldsymbol{\beta}}\right)$

(same applies to any other parameters $\boldsymbol{\alpha}$ we estimate as part of our model)

General purpose inference: several techniques

Maximum likelihood, ML. Assume a probability distribution for the response; derive the most appropriate estimator from it

ML simplest place to start, and can estimate most models,
including most of those we'll consider in this class

Markov Chain Monte Carlo, MCMC. Assume probability distributions for response and all parameters, and estimate a Bayesian model

MCMC almost always works, but requires more care and computing power

General Method of Moments, GMM. Determine the moments of the response (leave distribution vague); derive a model that "matches" them using covariates

GMM makes fewer assumptions than ML or MCMC.
Less efficient when distribution is known; more robust when not

We will focus on ML estimation, but GMM will come up,
and some examples will benefit from MCMC (or other simulation methods)

Model selection and interpretation

We leave details of MLE and GMM to other courses

Once we estimate least squares (or MLE or GMM),
our main concerns will be fitting and interpretation

Fitting: Choosing the right specification based on goodness of fit tests

Interpretation: Clearly showing the substantive meaning of a fitted model

Goodness of Fit

1. Why R^2 isn't usually useful

2. Why the standard error of the regression is useful

3. Out of sample fit and "over-fitting"

4. Goodness of fit for MLEs:
   Likelihood ratios and information criteria (AIC)

5. Cross-validation

Much ado about R^2

Many regression tables report R^2

R^2 is the proportion of variance in y explained by the regression model

$$R^2 = \frac{\text{Explained sum of squares}}{\text{Total sum of squares}} = 1 - \frac{\text{Residual sum of squares}}{\text{Total sum of squares}}$$

$$= \frac{\sum \hat{y}^2}{\sum y^2} = 1 - \frac{\sum \hat{\varepsilon}^2}{\sum y^2}$$

R^2 is bounded by 0 and 1

$R^2 = 0$: model has no explanatory power

$R^2 = 1$: model perfectly predicts every case without error

Old conventional wisdom in political science: higher R^2 is "better"

King ("How not to lie with statistics") on what R^2 does & doesn't say

- R^2 shows the *proportion of variance explained* by the covariates, compared to a model with no covariates

- R^2 indirectly reports the scatter around the regression line ($\hat{\sigma}^2$ directly reports the amount of scatter)

- R^2s comparable only for models with same observations and response variable (e.g., adding a covariate with missing values makes R^2 non-comparable)

- Maximizing R^2 is perverse as more covariates *always* raise R^2

- The most useful model is seldom the one with the highest R^2 (Uninteresting high R^2 models: y regressed on itself, vote choice regressed on vote intention)

- R^2 is not an "estimate," so it can't be significant or non-significant

- R^2 seldom of substantive interest, unlike $\hat{\sigma}^2$, the standard error of the regression

The standard error of the regression

$\hat{\sigma}$ is at least as useful to report as R^2

It tells us how much the fitted values, $\hat{y}$, miss the true y on average

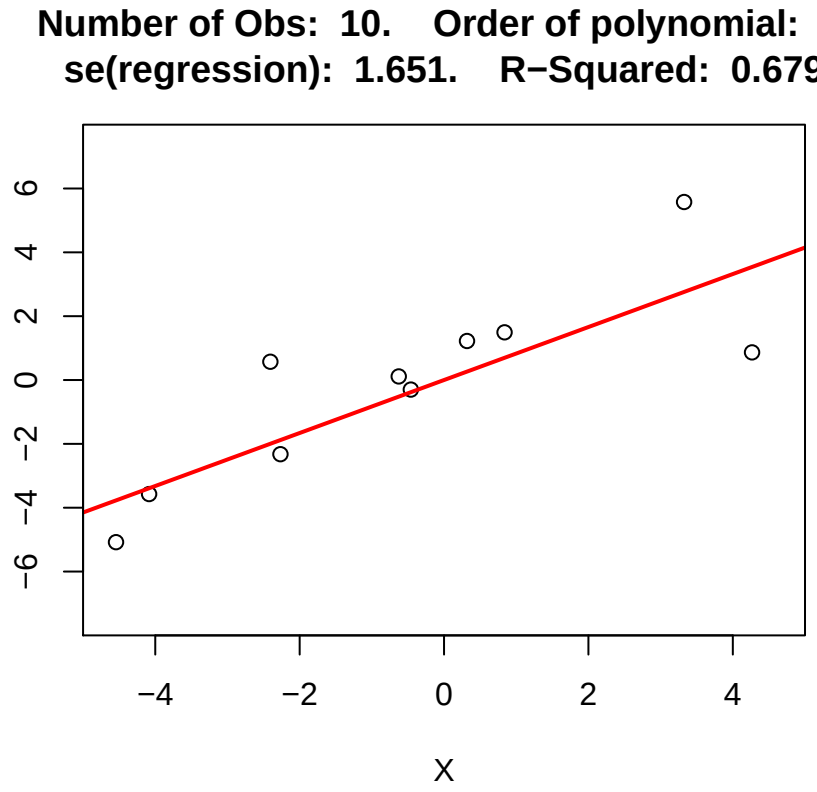
Because it's on the scale of y, $\hat{\sigma}$ is easy to work into a substantive conclusion

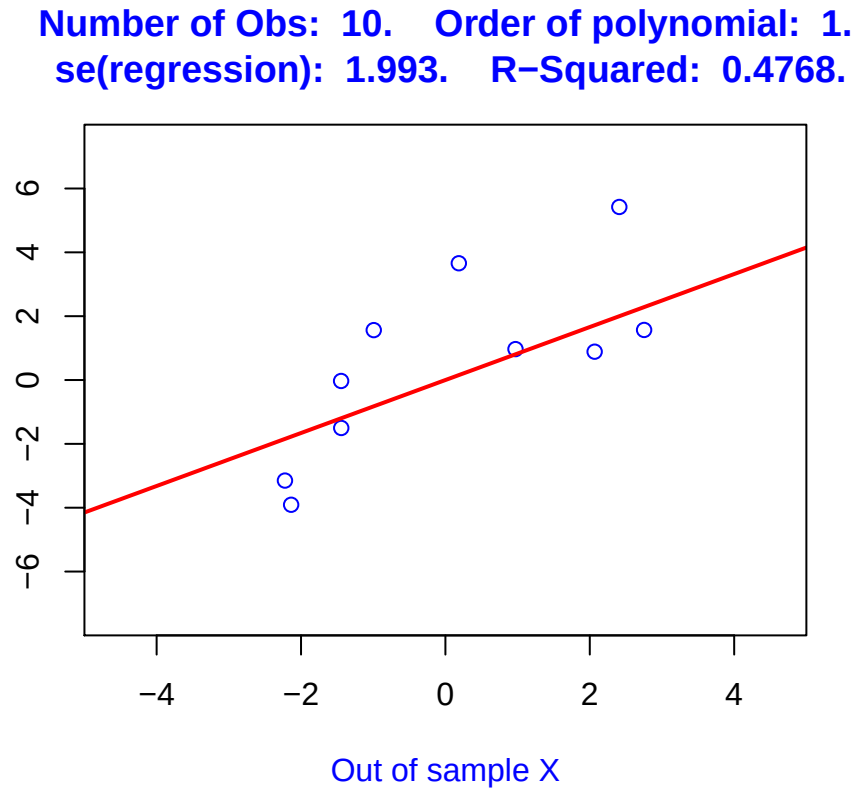
How is this a measure of goodness of fit?

As $\hat{\sigma}$ goes down, we make better predictions

$\hat{y}$ gets closer to the true y

So should we just minimize $\hat{\sigma}$?
Not necessarily – an example shows that we could easily "overfit" by choosing the model that minimizes in-sample error



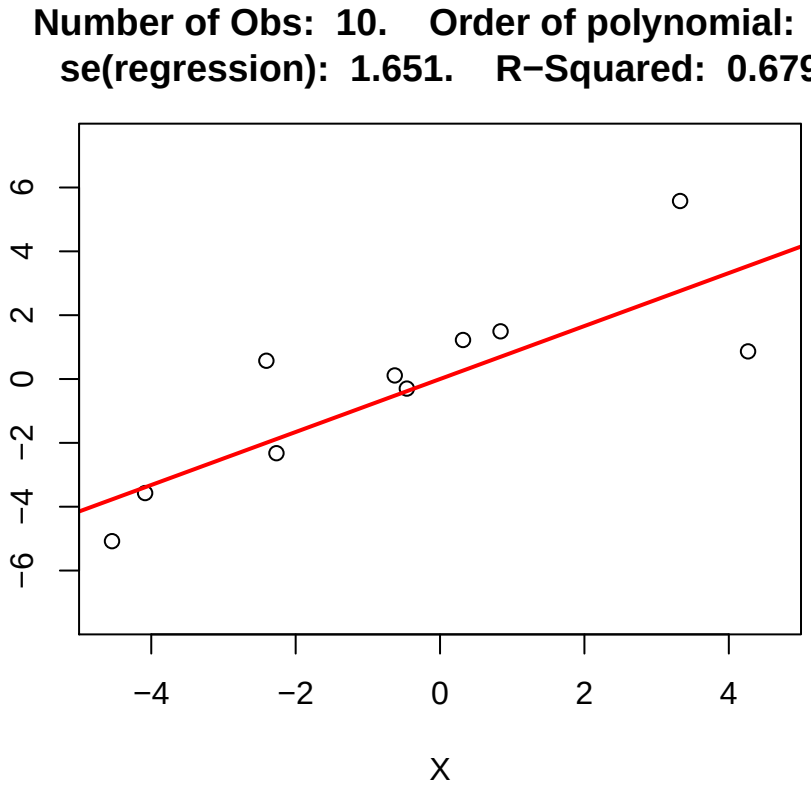


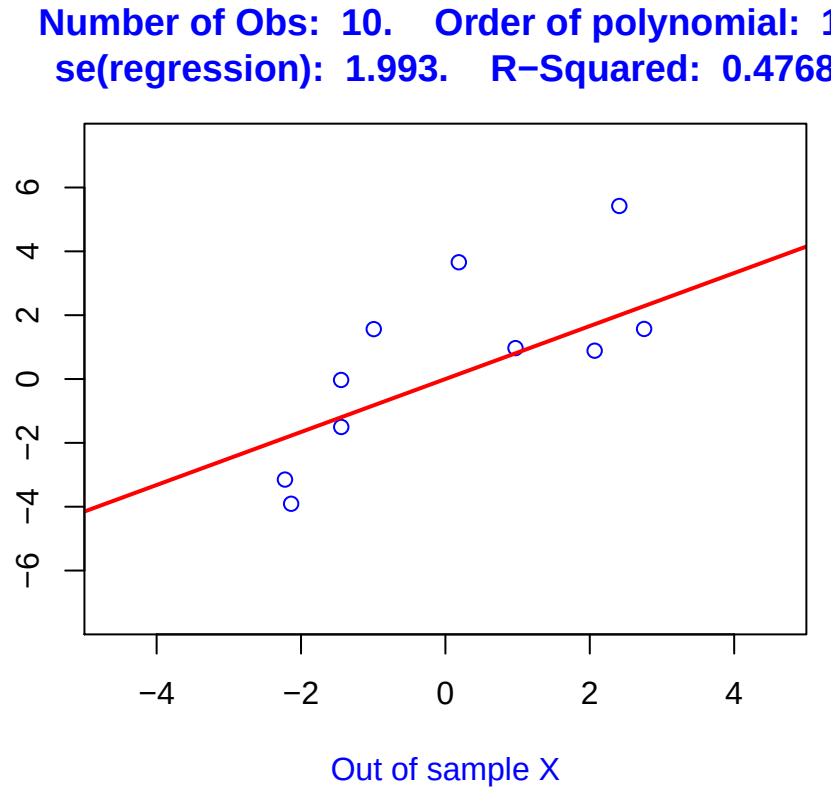
Polynomial overfitting experiment: generate 10 obs from the "true" model:

$$y = 0 + 1x + \varepsilon, \qquad \varepsilon \sim \mathrm{N}(0, 3)$$

and fit these data using different polynomials of x.

We will show the fit of the model to the original data on the left



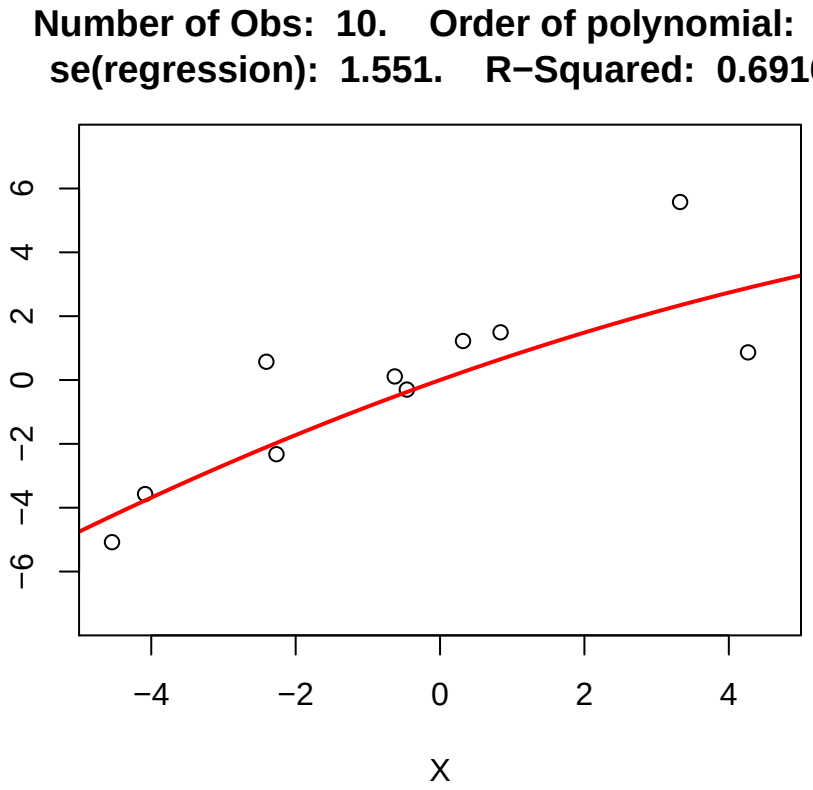


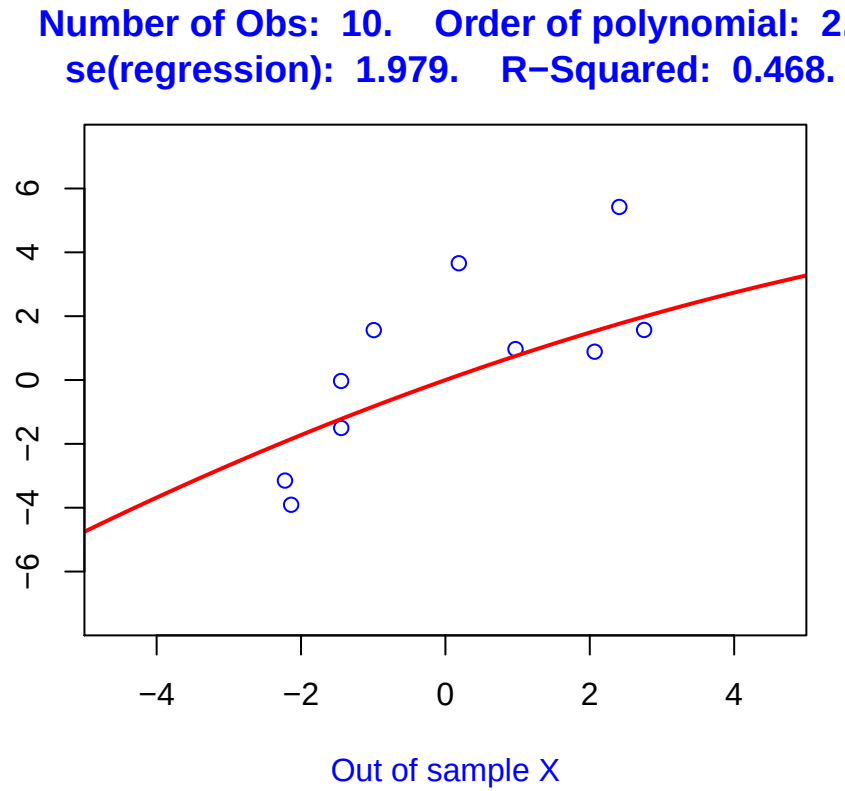
We will also draw new "out of sample data" (OOS) from the same true model:

$$y_{\text{OOS}} = 0 + 1x_{\text{OOS}} + \varepsilon_{\text{OOS}}, \qquad \varepsilon \sim \text{N}(0, 3)$$

and use the model as fitted on the original dataset to predict out of sample cases

We will show the fit of the old model to the out-of-sample data on the right

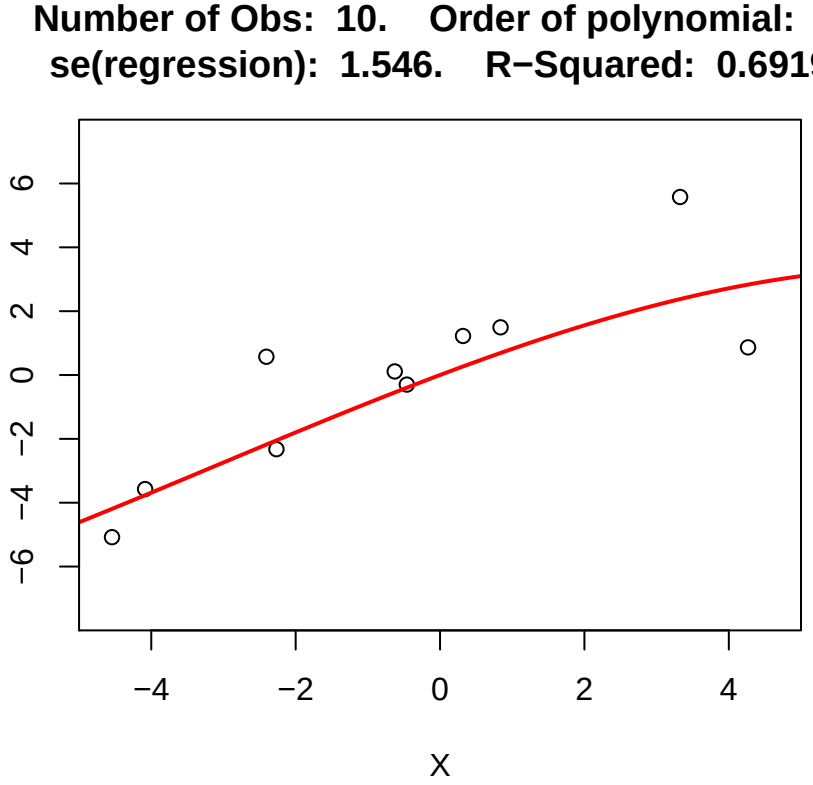


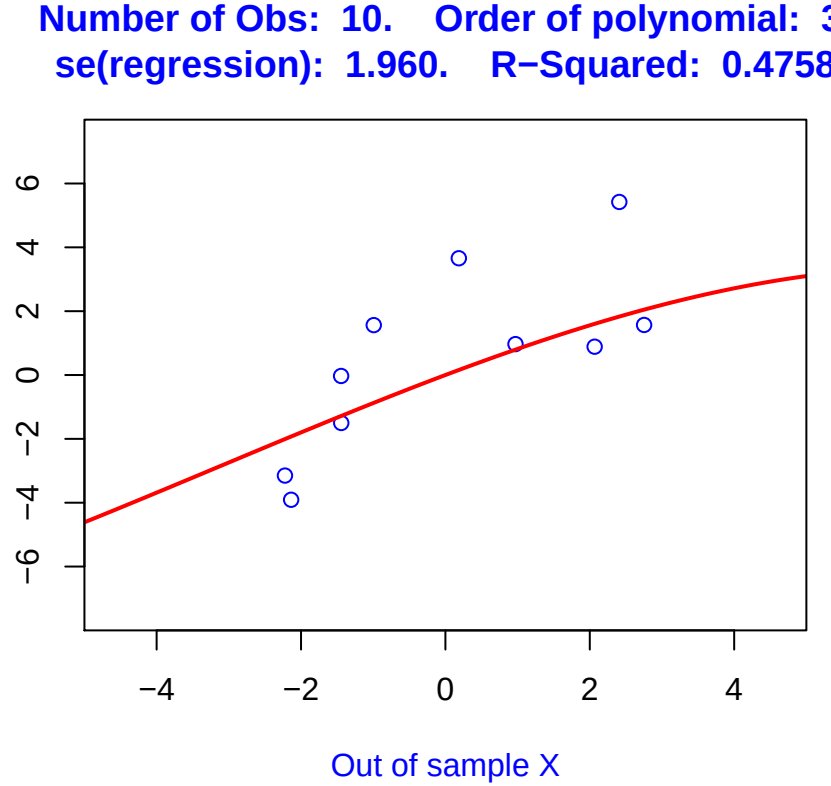


Above is the fit from a quadratic specification of x, ie, we estimated:

$$y = \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\varepsilon}$$

Note that we omit the constant for simplicity of exposition



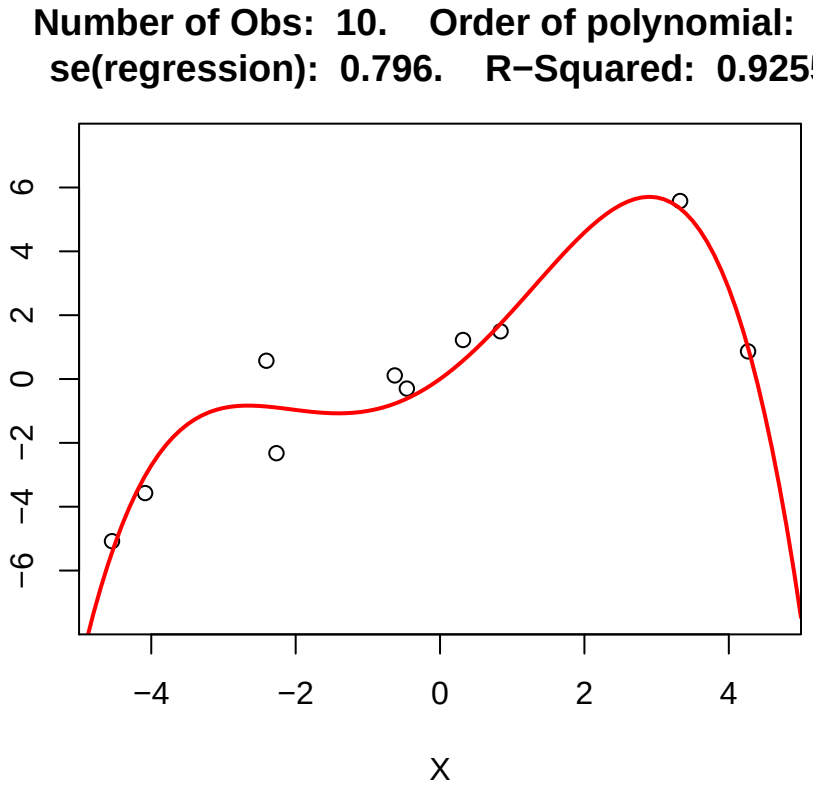


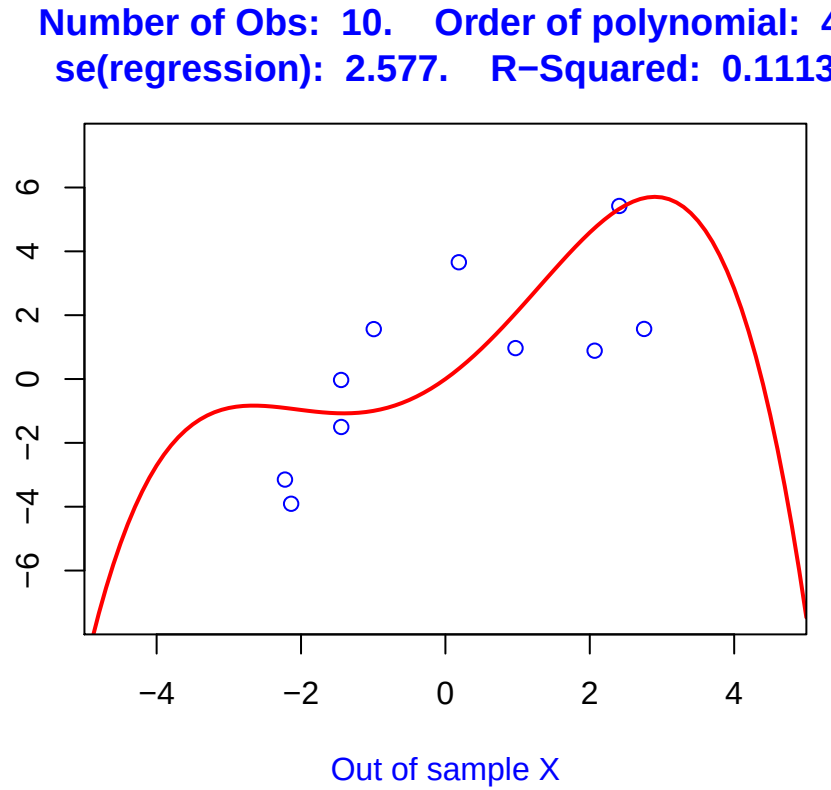
Above is the fit from a cubic specification of x; that is, we estimated:

$$y = \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \hat{\varepsilon}$$

How many polynomials can we add and still find $\hat{\beta}$?

What will happen to the fit in and out of sample as we add polynomials?

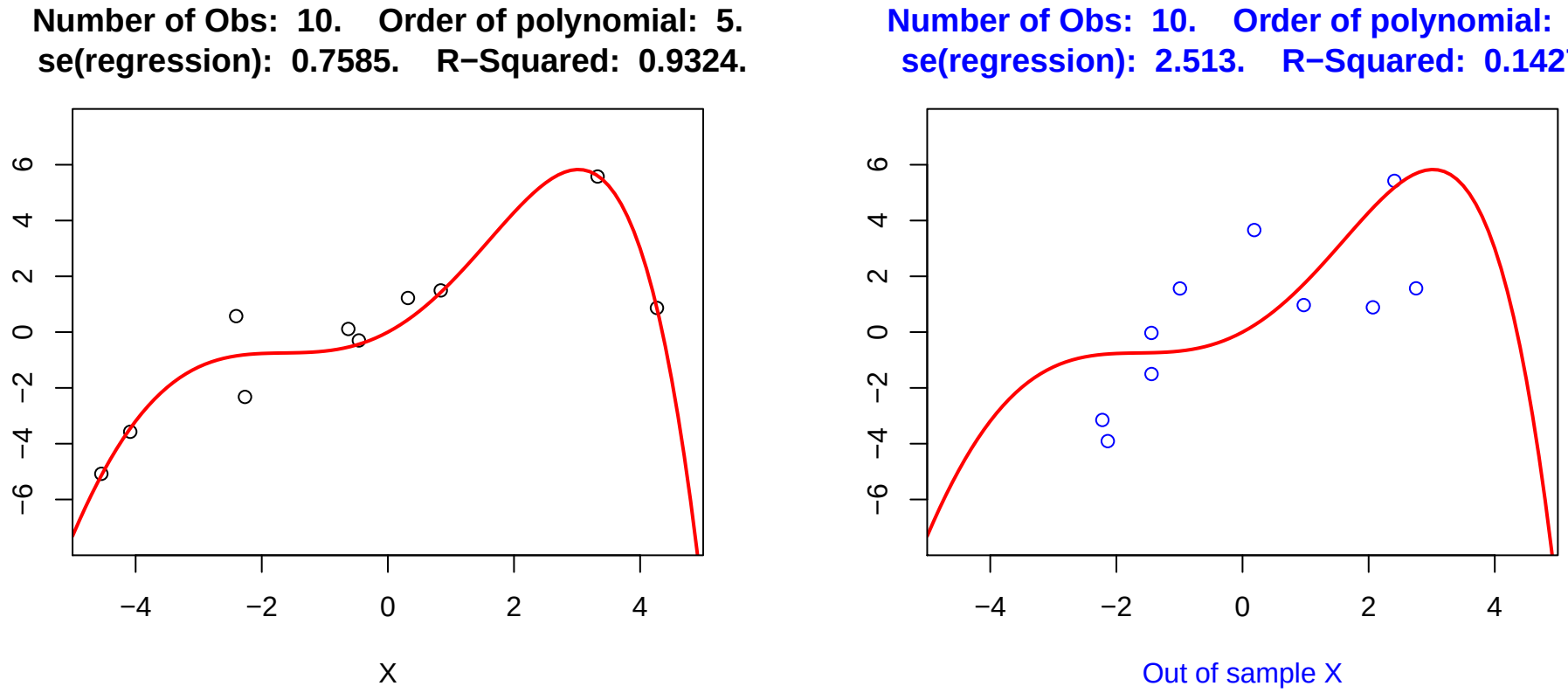




Above is the fit from a quartic specification of x; that is, we estimated:

$$y = \hat{\beta}_1 x + \hat{\beta}_2 x^2 + \hat{\beta}_3 x^3 + \hat{\beta}_4 x^4 + \hat{\varepsilon}$$

On the left, we see the model finds a curious non-linearity, by which low and high x suppress y, but middle values of x increase y. Do you trust this finding?

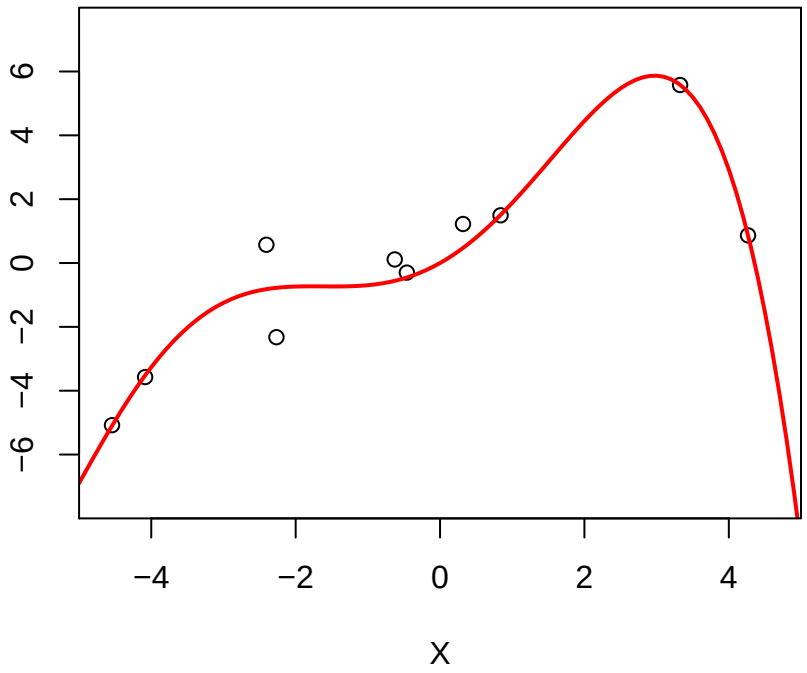


In small samples,
outliers can easily create the illusion of complex curves relating x and y

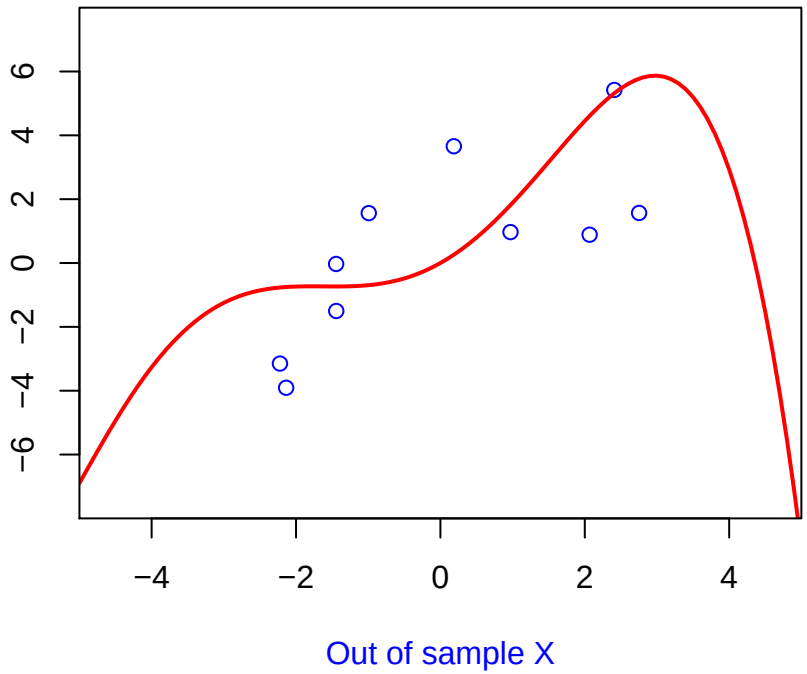
We need *a lot* of data spread across the range of x to discern if such curves are more than spurious

(And so we probably need a strong theory, too, to justify the data collection)

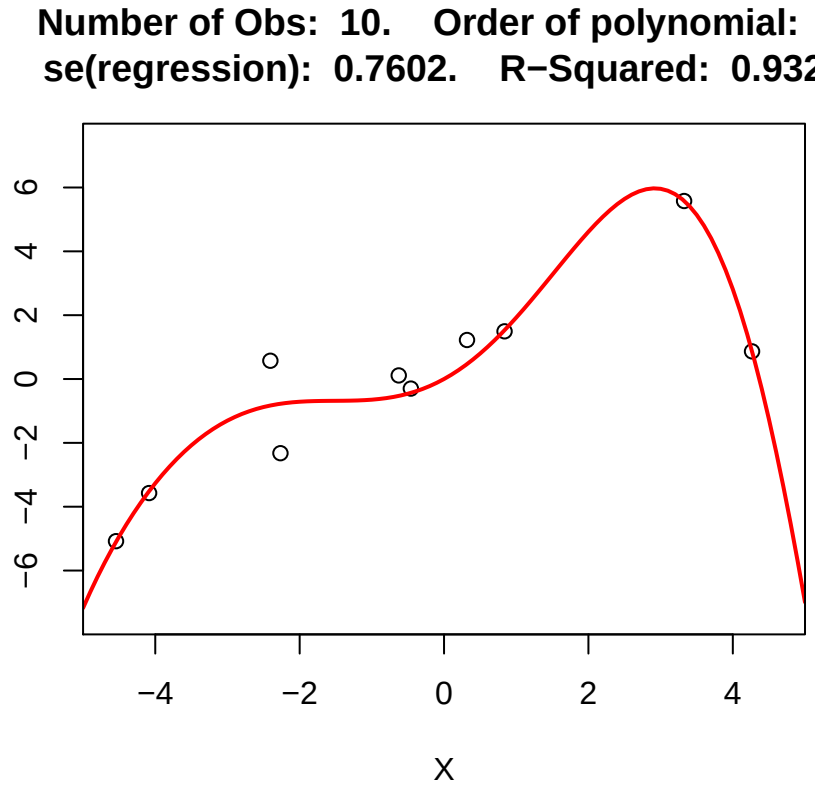
Number of Obs: 10. Order of polynomial: 6.
se(regression): 0.7591. R−Squared: 0.9326.

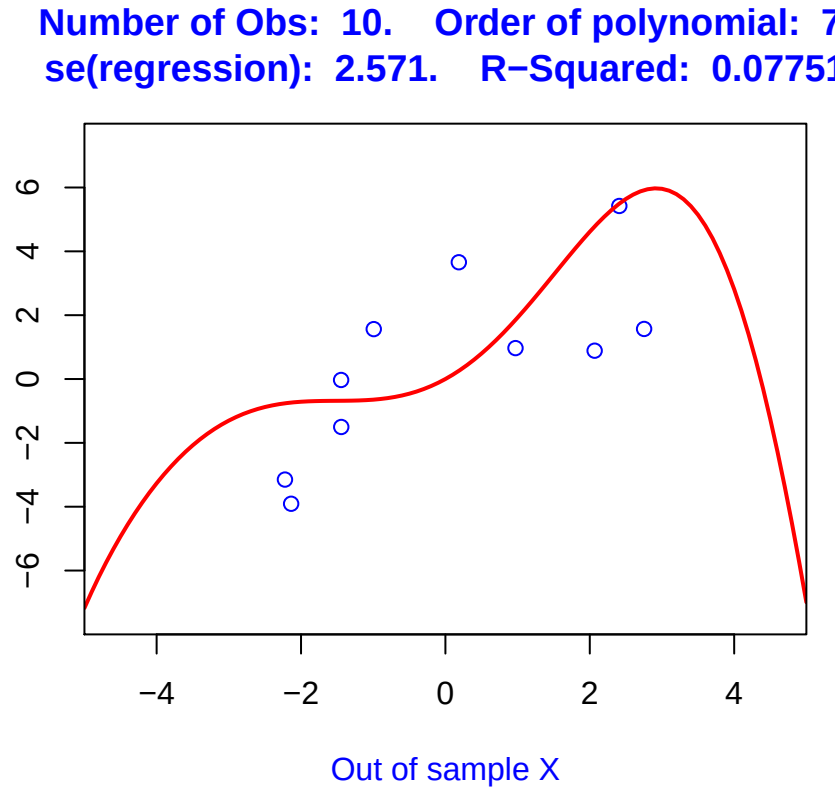


Number of Obs: 10. Order of polynomial: 6.
se(regression): 2.543. R−Squared: 0.1113.

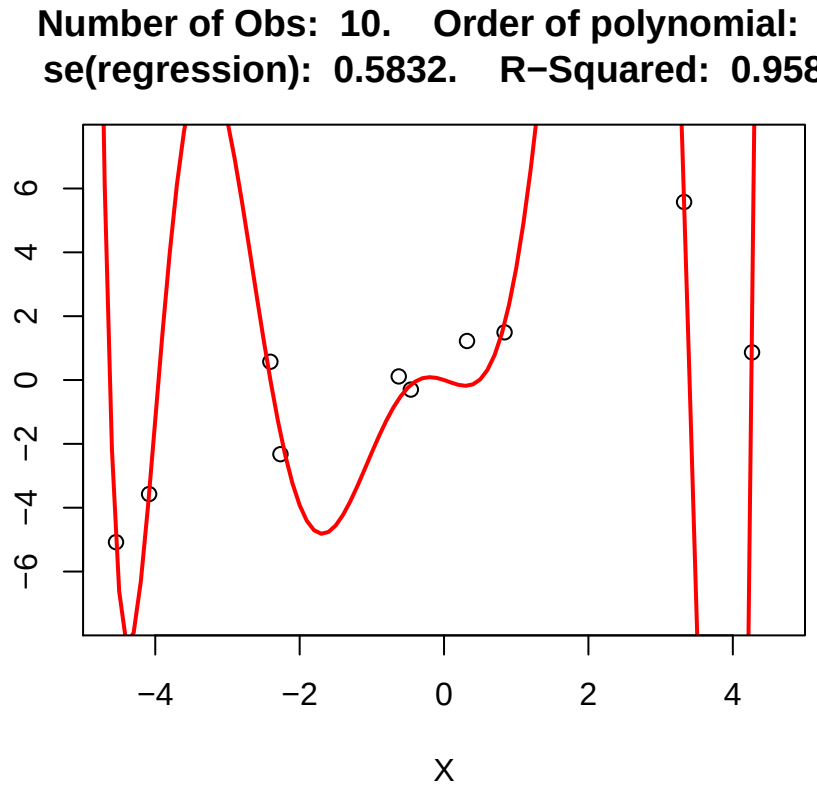


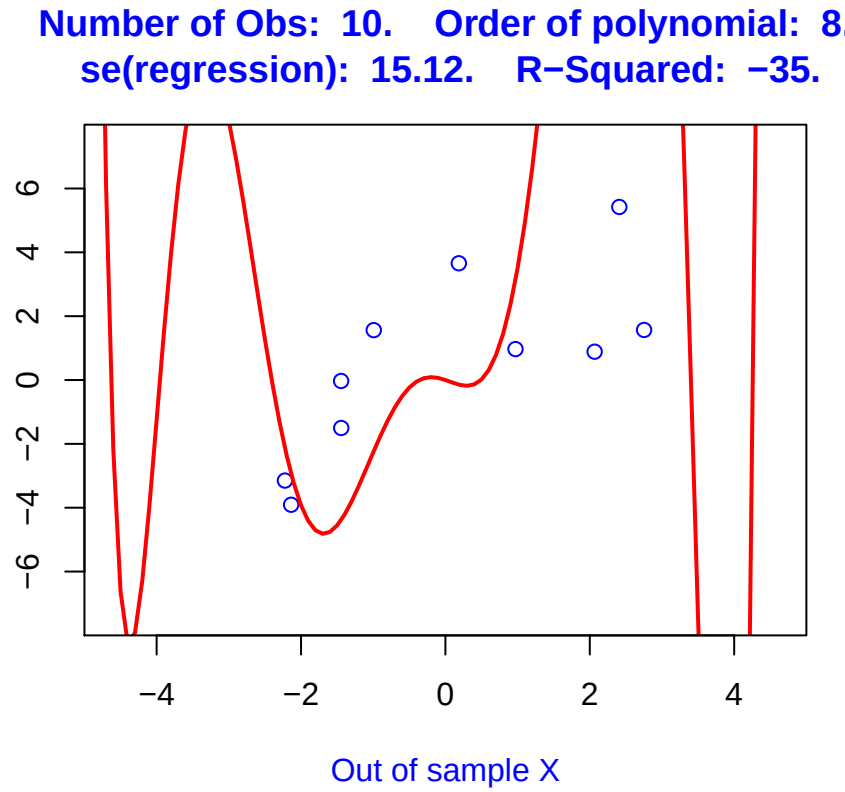
What happens as we approach a tenth-order polynomial?



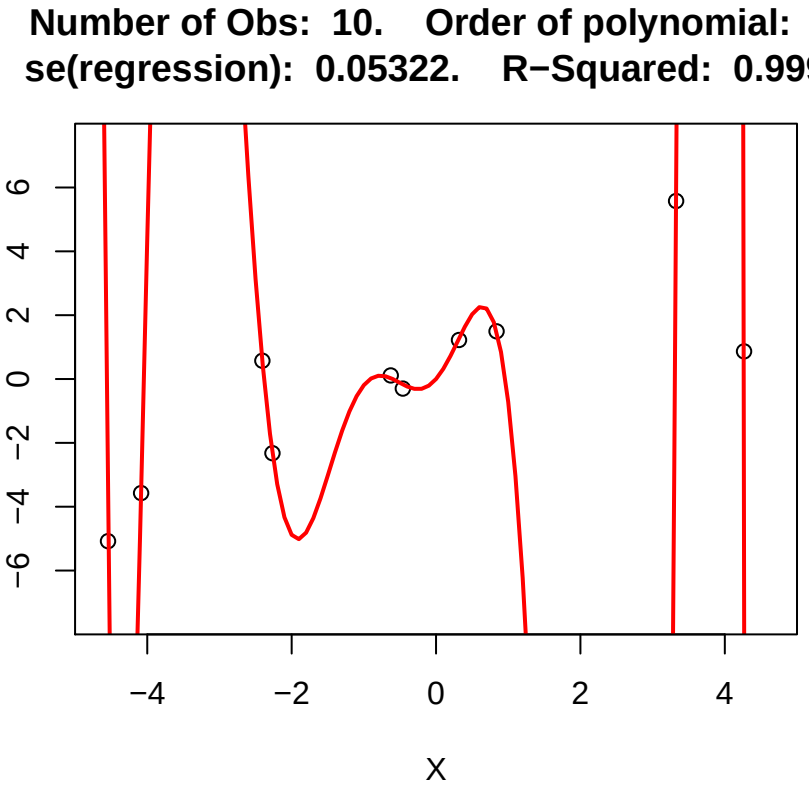


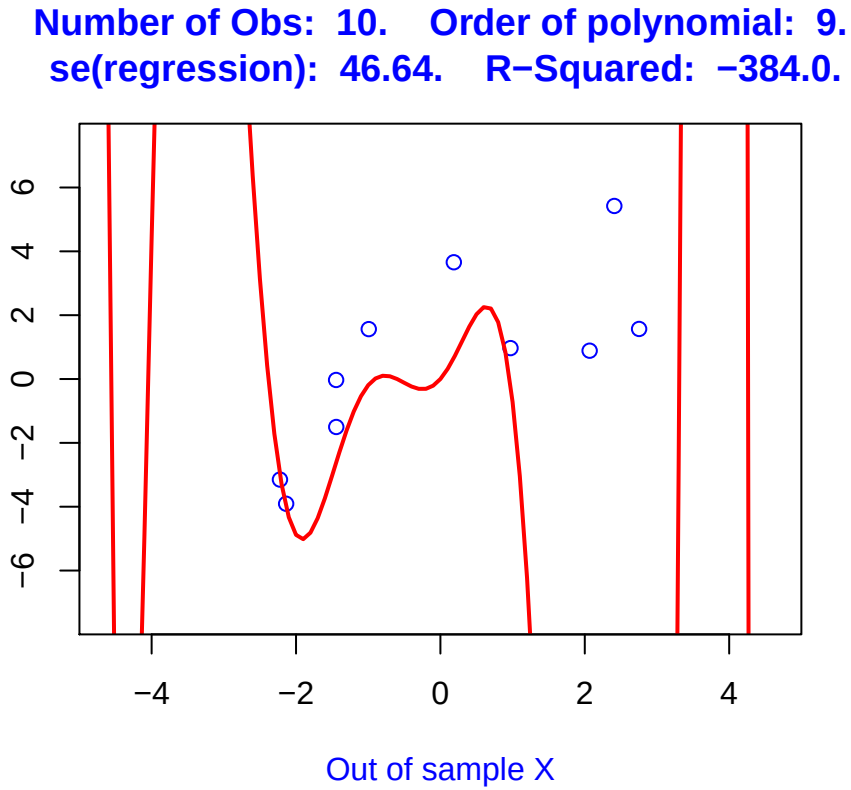
What happens as we approach a tenth-order polynomial?



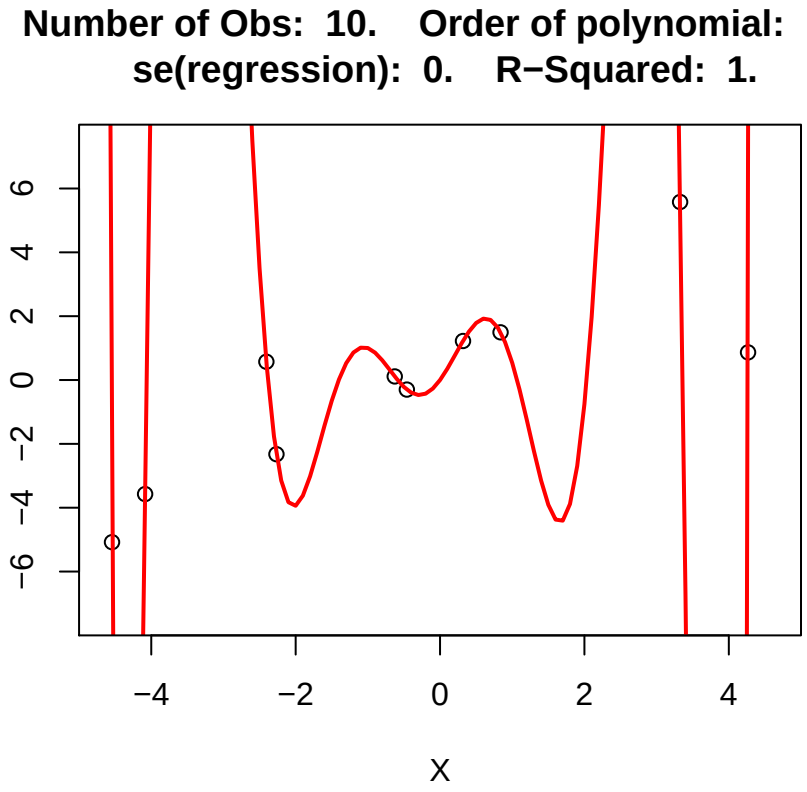


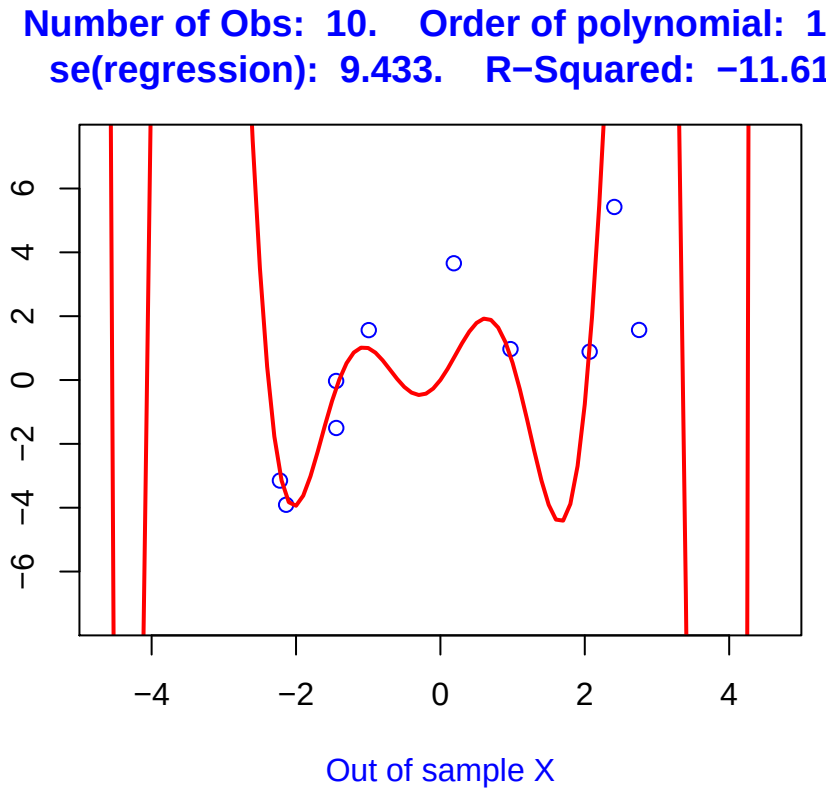
What happens as we approach a tenth-order polynomial?





What happens as we approach a tenth-order polynomial?

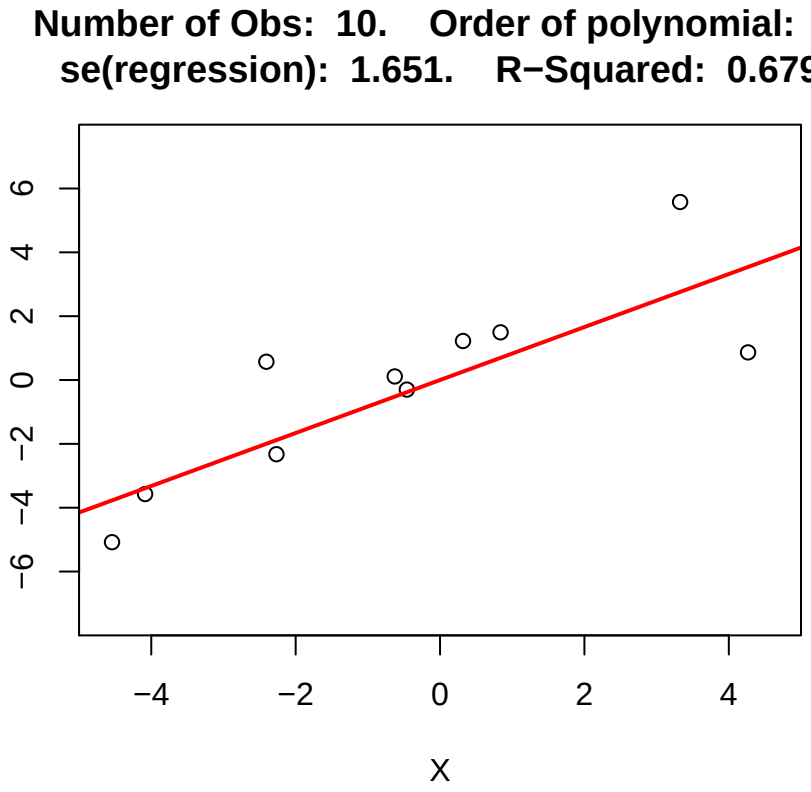


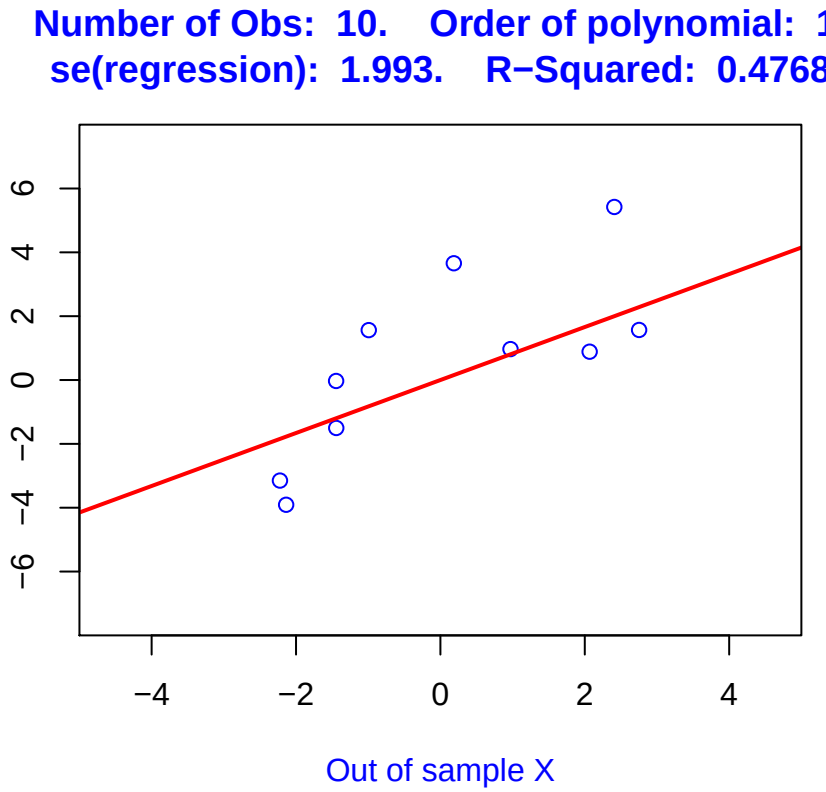


When the number of parameters in the model equals the number of observations, least squares is able to fit a line through every datapoint

This means the fit will be "perfect": no error

And out of sample, it will be completely useless,
and worse than guessing that y simply equals its sample mean in every case

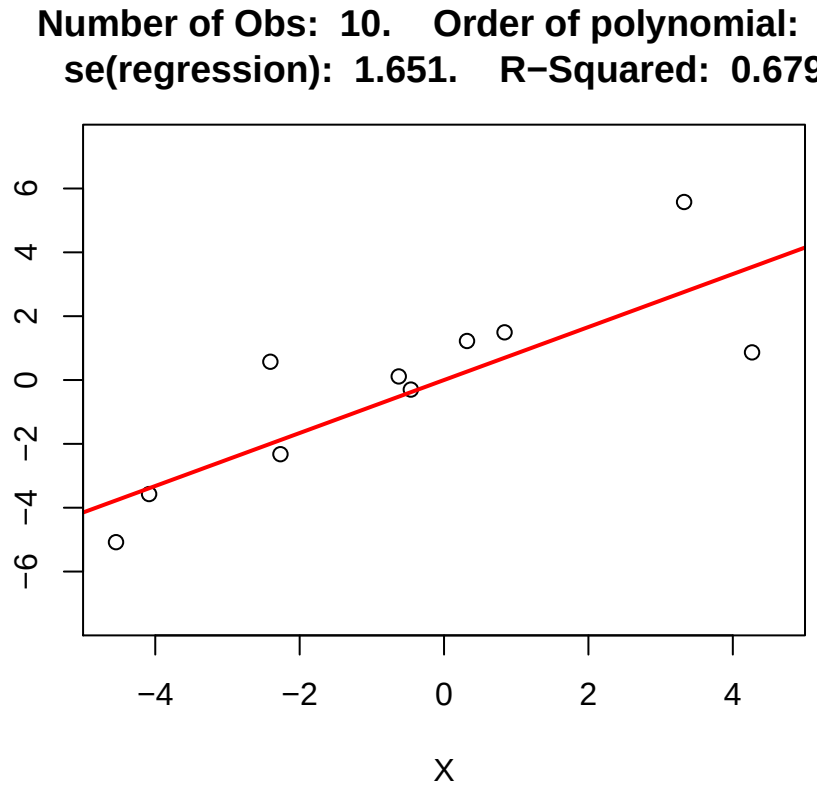


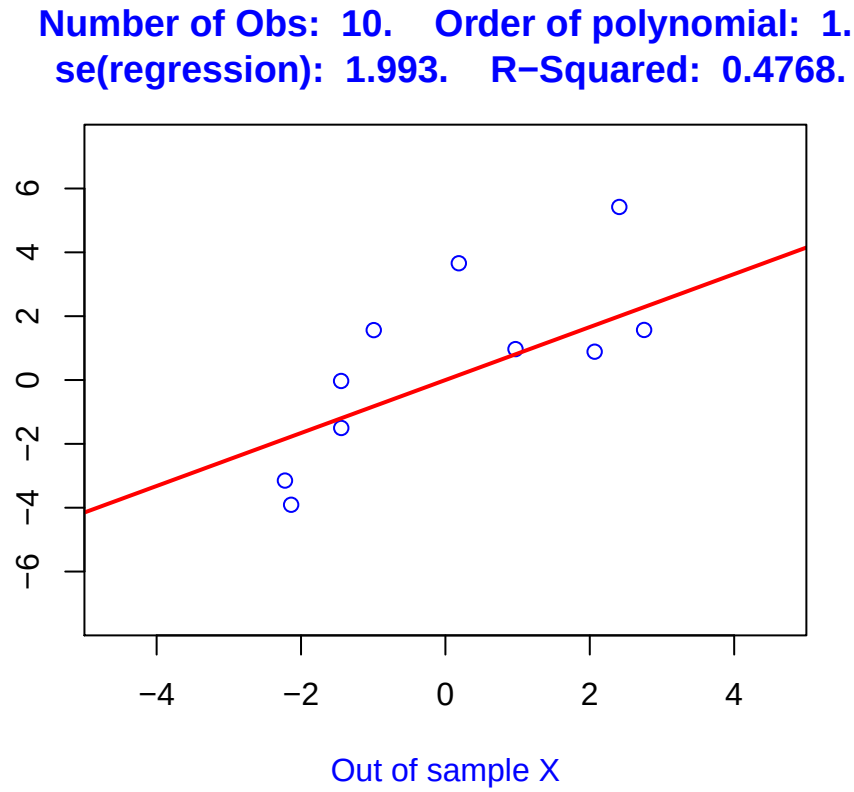


Two lessons:

1. Beware curve-fitting:

Significance tests may suggest adding polynomials,
but that isn't enough to justify their inclusion





2. Beware good fits in-sample unless they fit well out of sample, too

Every goodness of fit measure has an out-of-sample counterpart

Out of sample goodness of fit is *much more important* than in sample

Likelihood ratios

For maximum likelihood estimates (MLE),
many goodness of fit tests build on comparisons of "likelihoods" from different models

The likelihood is a (scaleless) measure that is higher for models that fit better (assuming the same outcome and observations)

This leads to the *likelihood ratio* test of nested models:

$$\mathrm{LR} = -2 \log \frac{L(\mathcal{M}_1)}{L(\mathcal{M}_2)}$$

$$\mathrm{LR} = 2 \left(\log L(\mathcal{M}_2) - \log L(\mathcal{M}_1) \right) \sim f_{\chi^2}(m)$$

where m is the number of restrictions placed on model 2 by model 1
(e.g., parameters held constant in model 1).

Obtaining a significant result (e.g., p less than a critical value)
is evidence for the encompassing model

Likelihood ratio tests require the same observations be used in each model

Information Criteria

Just like R^2 and s.e.r., LR tests tend to favor more complex models

This can lead to poor out of sample fit

Quick and dirty solution: *penalize* the likelihood for model complexity

Akaike's Information Criterion (AIC)

AIC applies a penalty for the number of parameters p in a model:

$$\text{AIC} = 2p - 2\log(\text{likelihood})$$

AIC should tend to be lower for models with better out-of-sample forecasting ability

Note that it is an *approximation*;
we haven't actually tested the model out of sample,
instead we just penalized in-sample fit

Other information criteria are available with different penalties
(e.g., the Bayesian Information Criterion, BIC)

More persuasive tests of fit

The most persuasive evidence of model fit is successful prediction

Suppose we estimate a model of US presidential approval ratings,
using data over 1950–1995

We could ask: How well does the model fit the 1950–1995 data?

Could also ask: How well does the model *predict* 1996–2005 data?

Out-of-sample tests are powerful checks against spurious findings

If we intend the model to apply to the out of sample data, this is a better test

The usual caveat applies:
the best fitting model is not necessarily the most interesting

But if our model fits as well to the out-of-sample data as the in-sample,
we are much more confident we found something real

Out of sample goodness of fit

How to do an out-of-sample test:

1. Fit the model on the training sample, $\{\mathbf{X}_{\text{training}}, \mathbf{y}_{\text{training}}\}$, obtain $\hat{\boldsymbol{\beta}}_{\text{training}}$

2. Use $\hat{\boldsymbol{\beta}}_{\text{training}}$ to calculate the fitted $\hat{\mathbf{y}}_{\text{test}}$ for the test sample, $\{\mathbf{X}_{\text{test}}, \mathbf{y}_{\text{test}}\}$

3. Compare fit for training sample to our ability to predict test sample.

   For example, compare $\hat{\sigma}_{\text{training}}$, the standard error of the residuals from the training regression, to the standard error of the "residuals" from the test predictions:

$$\text{std dev}(\mathbf{y}_{\text{test}} - \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{training}})$$

   This is the RMSE, or root mean squared (prediction) error

   We could simply choose the model with best out-of-sample RMSE;
   more work than computing AIC but more trustworthy

Cross-validation

What if we're short on data, and can't afford to sequester half of it for an out of sample GoF test? Not a problem: just use *cross-validation*

1. Select all but $1/k$th of the data, $\{\mathbf{y}_{\text{training}}, \mathbf{X}_{\text{training}}\}$; leave-one-out cross validation ($k = n$) is often good

2. Regress $\mathbf{y}_{\text{training}}$ on $\mathbf{X}_{\text{training}}$ to obtain β_{training}

3. Use $\hat{\boldsymbol{\beta}}_{\text{training}}$ and $\mathbf{X}_{\text{test}}$ to form predictions $\hat{\mathbf{y}}_{\text{test}}$

4. Repeat steps 1–4 k times, saving the prediction error for each test case predicted

5. Using all the collected test cases, compute the root mean squred prediction error:

$$\text{std dev}(\mathbf{y}_{\text{test}} - \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{\text{training}})$$

The best predictive model, according to cross-validation, will minimize this root mean squared predictive error. (You could also use the mean absolute error)

Substantive & statistical significance

Interpretation: Clearly showing the substantive meaning of a fitted model

A general warning about interpretation: don't over interpret p-values

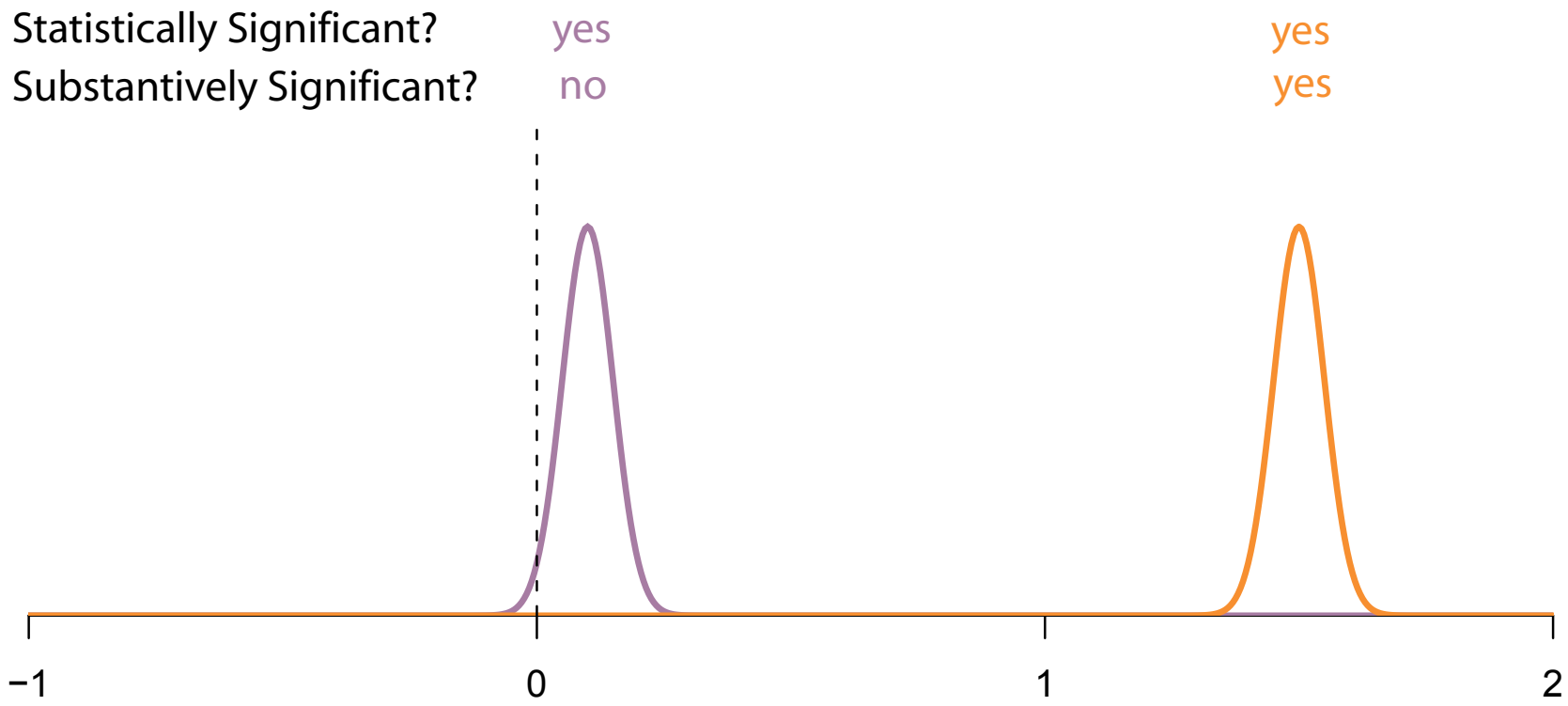
They only show *statistical* significance, not substantive significance

Statistical and substantive significance can interact

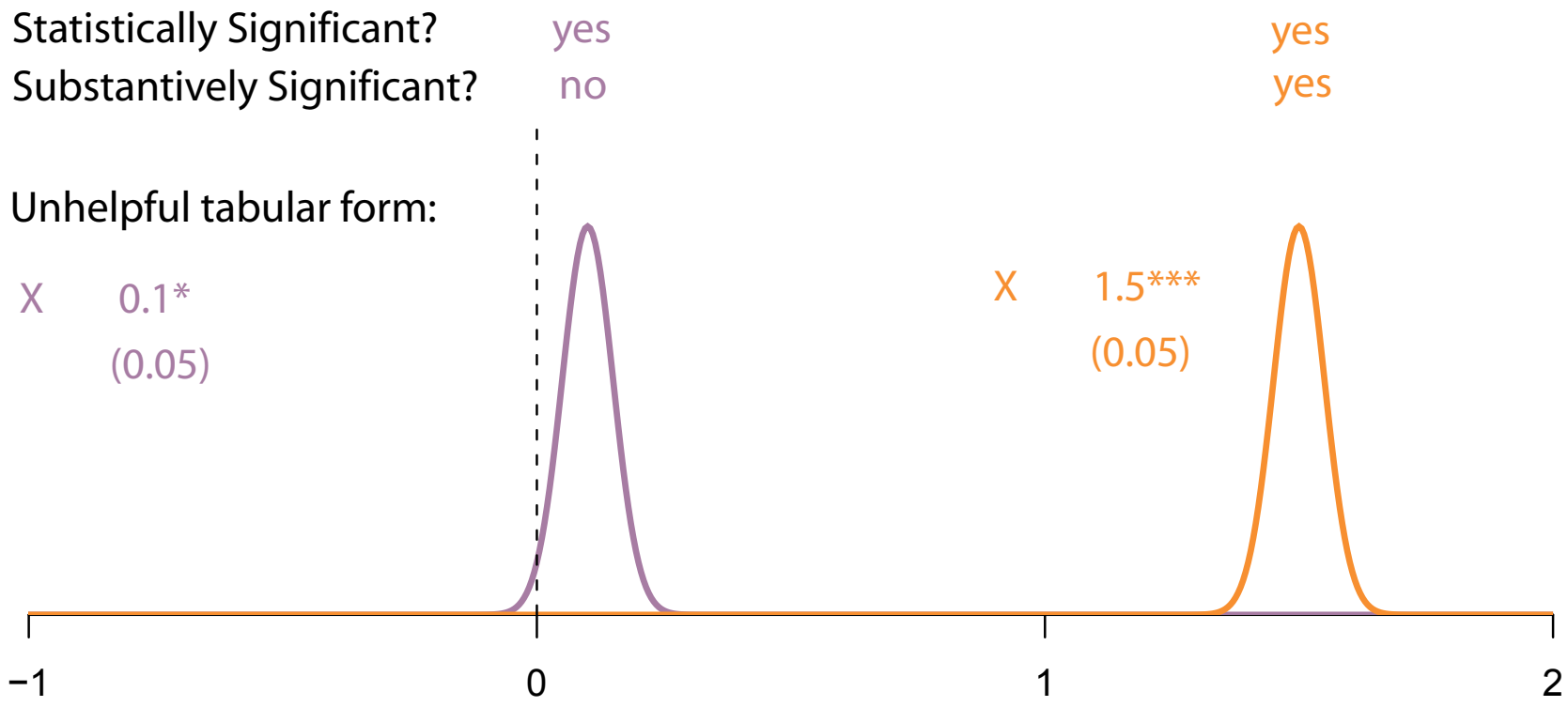
Sometimes non-significant results tell us something substantively;
other times, result with similar p have substantively different implications

A look at some hypothetical distributions of $\hat{\beta}_1$ helps frame the possibilities

Perils of stargazing



Perils of stargazing

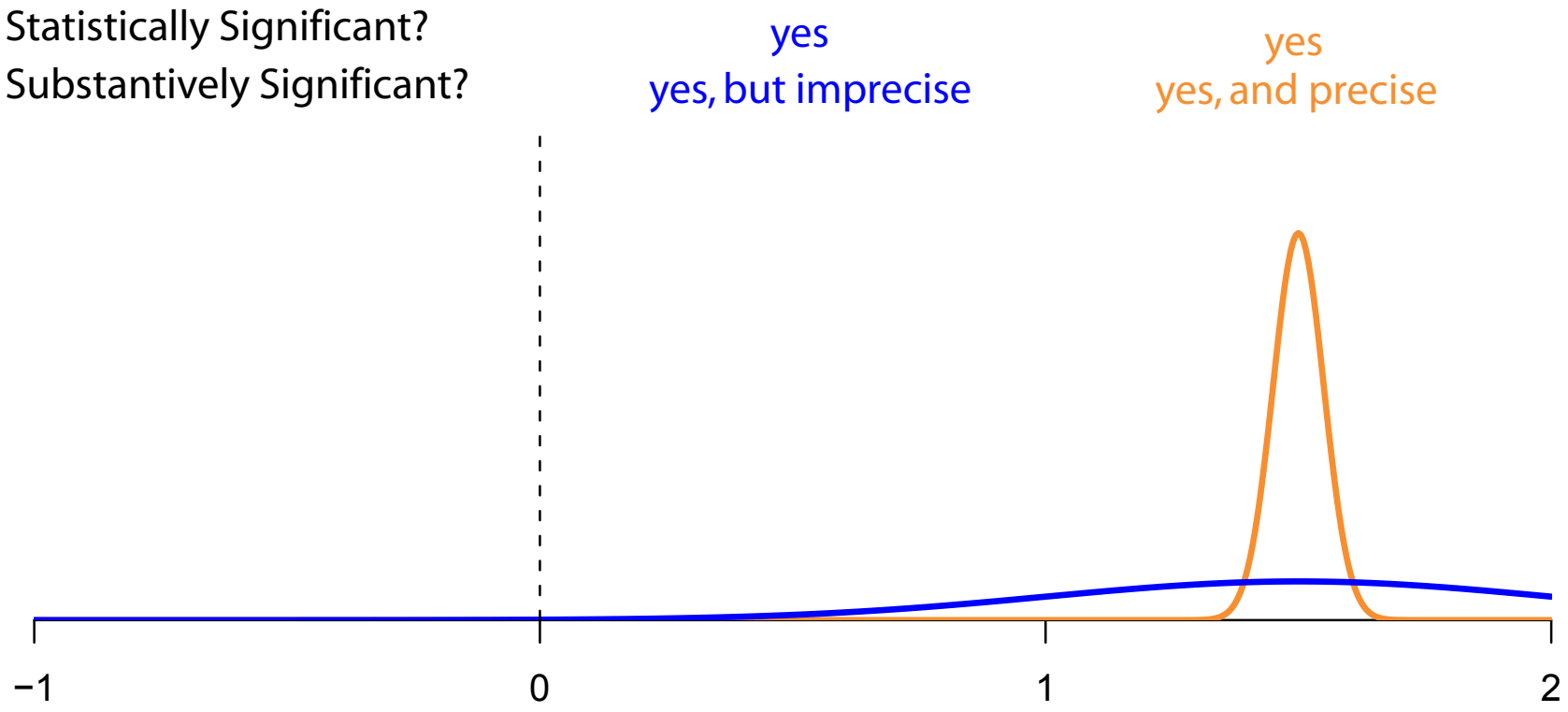


These estimated β's will both be starred in regression output.

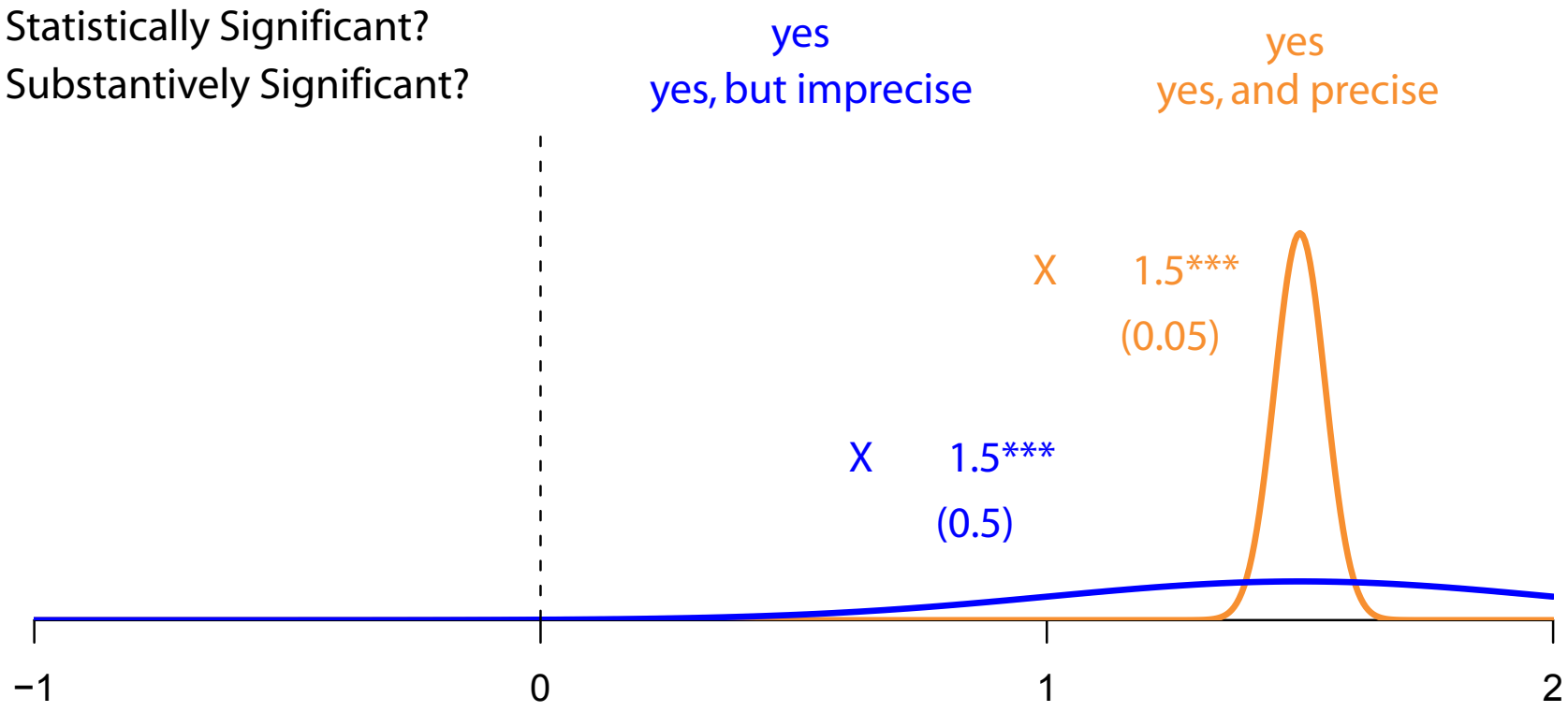
Often, only the estimate to the right will be significant in a substantive sense

The estimate on the left is a precise zero

Perils of stargazing



Perils of stargazing



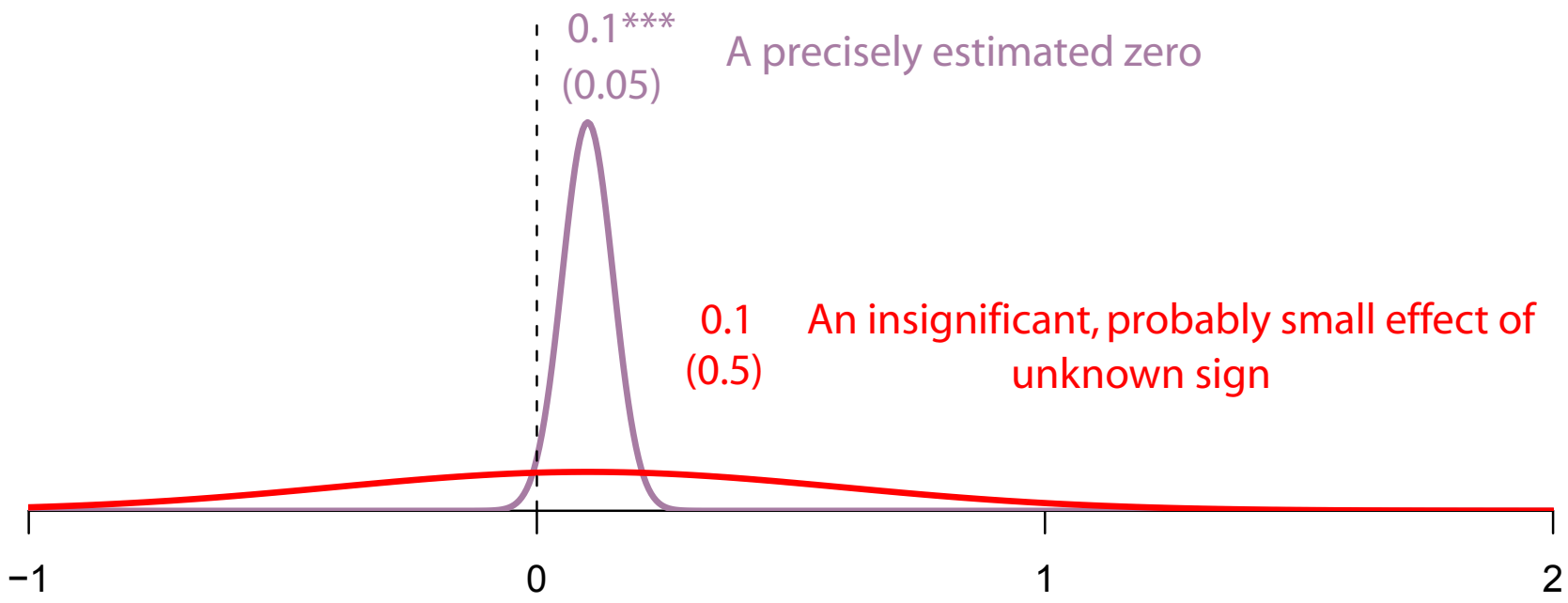
These estimated β's will both be heavily starred in regression output.

They are both substantively significant as well, with identical point estimates

But the orange curve is much more precisely estimated

The blue estimate may be much smaller or larger. Best shown with a CI

Perils of stargazing



How do you verify a null effect? Precise zeros (or "significant zeros")

Sometimes, researchers mistake the precise zero for a positive effect

Simulating quantities of interest

One way to avoid the perils of stargazing is to compute CIs around the substantive quantity of interest (e.g., the middle 95% region of the densities above)

Quantities of interest go beyond $\hat{\beta}$ to directly answer the research question:

- Predicted Values: $\tilde{y}|\mathbf{x}_c$
- Expected Values: $\mathrm{E}(y|\mathbf{x}_c)$
- First Differences: $\mathrm{E}(y|\mathbf{x}_{c2}) - \mathrm{E}(y|\mathbf{x}_{c1})$
- Risk Ratios: $\mathrm{E}(y|\mathbf{x}_{c2})/\mathrm{E}(y|\mathbf{x}_{c1})$
- Any function of the above

for some counterfactual $\mathbf{x}_c$'s.

But we also want the confidence intervals of these quantities,
which are tiresome to calculate analytically,

and we want a method that works for any model, not just regression.

Simulating quantities of interest

A generic simulation method (King, Tomz, Wittenberg 2000) for models like:

$$y_i \sim f(\mu_i, \alpha) \qquad \mu_i = g(\boldsymbol{\beta}, \mathbf{x}_i)$$

(for convenience, $\boldsymbol{\theta} = \mathrm{vec}(\boldsymbol{\beta}, \alpha)$)

This formulation includes many of the models in this class

And even for GMM models (which don't fit the above),
the following algorithms work

Four example algorithms for simulating QoIs

We want to understand the hypothetical behavior of y given counterfactual $\mathbf{x}_c$

Assume we have point estimates $\hat{\boldsymbol{\beta}}$ and their variance covariance matrix, $\mathrm{V}\left(\hat{\boldsymbol{\beta}}\right)$

All simulations are uncertain due to *estimation uncertainty* of model parameters

Approach 1: **predicted values** $\tilde{y}|\mathbf{x}_c$,
which *also* include the uncertainty due to shocks, ε (so-called *fundamental* uncertainty)

For this approach, we also need the estimated variance of these shocks, $\hat{\sigma}^2$

Approach 2: **expected values** $\hat{y}|\mathbf{x}_c$,
which average over the expected shocks

Expected values average over shocks to reveal the average level of y given the counterfactual covariates

Four example algorithms for simulating QoIs

We want to understand the hypothetical behavior of y given counterfactual $\mathbf{x}_c$

Assume we have point estimates $\hat{\boldsymbol{\beta}}$ and their variance covariance matrix, $\mathrm{V}\left(\hat{\boldsymbol{\beta}}\right)$

All simulations are uncertain due to *estimation uncertainty* of model parameters

Approach 3: **first differences** $\hat{y}|\mathbf{x}_{\text{post}} - \hat{y}|\mathbf{x}_{\text{pre}}$

The expected difference in the outcome given a difference in covariates

Approach 4: **risk ratios** $\hat{y}|\mathbf{x}_{\text{post}} / \hat{y}|\mathbf{x}_{\text{pre}}$

The factor change in the outcome given a difference in covariates;
also known as a *relative risk*

Counterfactual forecasting: Predicted Values

1. Choose hypothetical $\mathbf{x}_c$'s.

2. Draw a vector of simulated parameters from their asymptotic distribution: $\tilde{\boldsymbol{\beta}} \sim \mathcal{MVN}\left(\hat{\boldsymbol{\beta}}, \operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right)\right)$.

3. Draw a new random shock $\tilde{\varepsilon} \sim \mathcal{N}(0, \hat{\sigma}^2)$.

4. Calculate one simulated predicted value of y using

$$\tilde{y}|\mathbf{x}_c, \tilde{\boldsymbol{\beta}} = \mathbf{x}_c\tilde{\boldsymbol{\beta}} + \tilde{\varepsilon}.$$

5. Repeat steps 2–3 `sims` times to construct `sims` simulated predicted values. Summarize these predicted values by means and quantiles (prediction intervals).

Counterfactual forecasting: Expected Values

1. Choose hypothetical $\mathbf{x}_c$'s.

2. Draw a vector of simulated parameters from their asymptotic distribution: $\tilde{\boldsymbol{\beta}} \sim \mathcal{MVN}\left(\hat{\boldsymbol{\beta}}, \operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right)\right)$.

3. Calculate one simulated expected value of y using

$$\mathbb{E}\left(\tilde{y}|\tilde{\boldsymbol{\beta}}, \mathrm{x}_c\right) = \mathbf{x}_c\tilde{\boldsymbol{\beta}}.$$

4. Repeat steps 2–3 `sims` times to construct `sims` simulated expected values. Summarize these predicted values by means and quantiles (confidence intervals).

Counterfactual forecasting: First Differences

1. Choose hypothetical $\mathbf{x}_{\text{pre}}$'s and x_{post}'s.

2. Draw a vector of simulated parameters from their asymptotic distribution: $\tilde{\boldsymbol{\beta}} \sim \mathcal{MVN}\left(\hat{\boldsymbol{\beta}}, \operatorname{Var}\left(\hat{\boldsymbol{\beta}}\right)\right)$.

3. Calculate one simulated first difference using

$$\mathbb{E}\left(\tilde{y}_{\text{post}} - \tilde{y}_{\text{pre}} | \tilde{\boldsymbol{\beta}}, \mathrm{x}_{\text{pre}}, \mathrm{x}_{\text{post}}\right) = \mathbf{x}_{\text{post}}\tilde{\boldsymbol{\beta}} - \mathbf{x}_{\text{pre}}\tilde{\boldsymbol{\beta}}.$$

4. Repeat steps 2–3 `sims` times to construct `sims` simulated first differences. Summarize these predicted values by means and quantiles (confidence intervals).

Counterfactual forecasting: Risk Ratios

1. Choose hypothetical $\mathbf{x}_{\text{pre}}$'s and x_{post}'s.

2. Draw a vector of simulated parameters from their asymptotic distribution: $\tilde{\boldsymbol{\beta}} \sim \mathcal{MVN}\left(\hat{\boldsymbol{\beta}}, \mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)\right)$.

3. Calculate one simulated risk ratios using

$$\mathbb{E}\left(\tilde{y}_{\text{post}}/\tilde{y}_{\text{pre}}|\tilde{\boldsymbol{\beta}}, \mathrm{x}_{\text{pre}}, \mathrm{x}_{\text{post}}\right) = \mathbf{x}_{\text{post}}\tilde{\boldsymbol{\beta}}/\mathbf{x}_{\text{pre}}\tilde{\boldsymbol{\beta}}.$$

4. Repeat steps 2–3 `sims` times to construct `sims` simulated risk ratios. Summarize these predicted values by means and quantiles (confidence intervals).

Forward to time series and panel data

These slides reviewed the basics of linear modeling:

1. Notation and estimation of linear regression

2. When linear regression will have good properties,
   and when it will produce poor estimates

3. How to select models using in-sample and out-of-sample criteria

4. How to interpret linear models in substantively clear ways,
   without being mislead by stargazing

Going forward, we will develop tools to solve these problems
for data that occur over time and across units and time