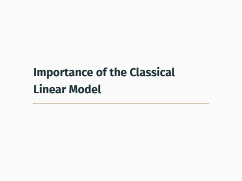
Importance of the Classical Linear Model

2020-07-22

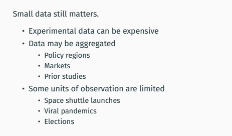
└─Importance of the Classical Linear Model



Importance of the Classical Linear Model

Small data still matters.

- Experimental data can be expensive
- Data may be aggregated
  - Policy regions
  - Markets
  - Prior studies
- Some units of observation are limited
  - Space shuttle launches
  - Viral pandemics
  - Elections



Small data still matters.

- Experimental data can be expensive
- Data may be aggregated
  - Policy regions
  - Markets
  - Prior studies
- Some units of observation are limited
  - Space shuttle launches
  - Viral pandemics
  - Elections

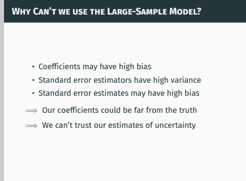
1. Why are we devoting this unit to the classical linear model?
2. Well, there's one big reason and a bunch of small ones.
3. The big reason is that, for all the talk about big data, small data still matters
4. As a data scientist, you are very likely to work with small data at some point in your career. There's a few reasons for that...

Why Can’t we use the Large-Sample Model?

- Coefficients may have high bias
- Standard error estimators have high variance
- Standard error estimates may have high bias

$\Longrightarrow$ Our coefficients could be far from the truth

$\Longrightarrow$ We can’t trust our estimates of uncertainty


Why Can’t we use the Large-Sample Model?

- Coefficients may have high bias
- Standard error estimators have high variance
- Standard error estimates may have high bias

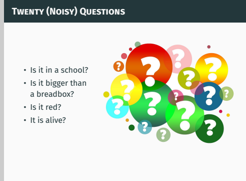
$\Longrightarrow$ Our coefficients could be far from the truth

$\Longrightarrow$ We can’t trust our estimates of uncertainty

1. First, let’s clearly say why we can’t just use the large-sample model for small data
2. In the large sample model, we rely on consistency and asymptotic normality, but those require a large n.
3. For small n, our coefficients naturally have higher variance, but they may also have high bias
4. That’s a problem because standard errors only account for variance, they don’t recognize bias
5. Furthermore, we have to estimate our standard errors, but those use consistent estimators.
6. In small data, our standard error estimates have high variance, and they might be very biased as well.
7. So we’re off but our instruments for understanding how

Twenty (Noisy) Questions

- Is it in a school?
- Is it bigger than a breadbox?
- Is it red?
- It is alive?



Twenty (Noisy) Questions

- Is it in a school?
- Is it bigger than a breadbox?
- Is it red?
- It is alive?

1. Imagine a noisy game of 20 questions.
2. I'm thinking of a secret object. You're trying to guess it by asking yes or no questions.
3. This is a noisy game, so my answers can be wrong with some probability - say 10%.
4. To represent small data, imagine that you get just 5 questions.
5. To represent biased coefficients, say that I can strategically choose when to give the wrong answer to try to direct you to a false target.
6. To represent biased standard errors, imagine that you have no idea what fraction of the time I'm allowed to lie.
7. The game doesn't sound like very much fun, does it? This

More assumptions

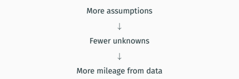
↓

Fewer unknowns

↓

More mileage from data

└Importance of the Classical Linear Model


More assumptions

↓

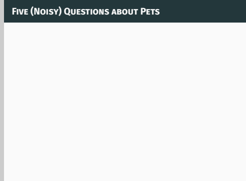
Fewer unknowns

↓

More mileage from data

1. So how do fix this? The super short answer is this: we need more assumptions.
2. More assumptions means fewer unknowns. Which means that we can get more mileage out of our data.

Five (Noisy) Questions about Pets

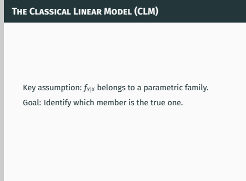


1. Back to the twenty questions game. You still get just 5 questions. But for this game, you get some extra assumptions at the start of the game. You assume that I'm thinking of a pet. And you assume that I can only answer incorrectly randomly, I can't be strategic.
2. If those assumptions are correct, you have a much better chance of winning this game. Or at least narrowing down the answer to a small set of possibilities.

The Classical Linear Model (CLM)

Key assumption: $f_{Y|X}$ belongs to a parametric family.

Goal: Identify which member is the true one.



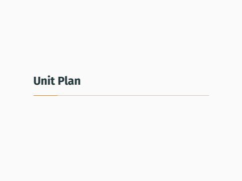
The Classical Linear Model (CLM)

Key assumption: $f_{Y|X}$ belongs to a parametric family.

Goal: Identify which member is the true one.

1. The enlarged set of assumptions we're talking about is known as the Classical Linear Model.
2. We're going to assume that the conditional distribution of Y belongs to a parametric family.
3. Then our job is much smaller. all we need to do is figure out which member of the family is the true one.
4. The CLM Is extremely popular. It's the traditional starting point for linear regression, and the centerpiece of most textbooks in statistics, econometrics, and machine learning.
5. It's also a basis for other statistical models. So it's very important to know this model. let's get started..

Unit Plan


Unit Plan

Plan for the Week

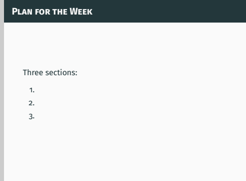
Three sections:

1.
2.
3.

2020-07-22

└─Unit Plan

└─Plan for the Week



Plan for the Week

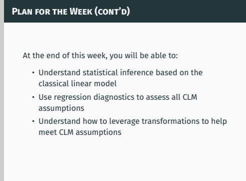
Three sections:

1.
2.
3.

Plan for the Week (cont'd)

At the end of this week, you will be able to:

- Understand statistical inference based on the classical linear model
- Use regression diagnostics to assess all CLM assumptions
- Understand how to leverage transformations to help meet CLM assumptions



Plan for the Week (cont'd)

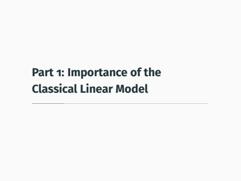
At the end of this week, you will be able to:

- Understand statistical inference based on the classical linear model
- Use regression diagnostics to assess all CLM assumptions
- Understand how to leverage transformations to help meet CLM assumptions

Part 1: Importance of the Classical Linear Model

2020-07-22

Part 1: Importance of the Classical Linear Model



Part 1: Importance of the Classical Linear Model

Reading: The CLM Assumptions

2020-07-22

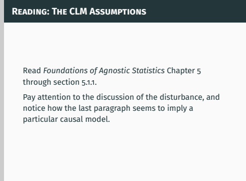
└Reading: The CLM Assumptions

Reading: The CLM Assumptions

Reading: The CLM Assumptions

Read *Foundations of Agnostic Statistics* Chapter 5 through section 5.1.1.

Pay attention to the discussion of the disturbance, and notice how the last paragraph seems to imply a particular causal model.



Reading: The CLM Assumptions

Read *Foundations of Agnostic Statistics* Chapter 5 through section 5.1.1.

Pay attention to the discussion of the disturbance, and notice how the last paragraph seems to imply a particular causal model.

The CLM Assumptions

2020-07-22

└─The CLM Assumptions

The CLM Assumptions

The CLM Assumptions

1. The book presents a very compact version of the CLM, and it's 100% correct. But we're going to break it apart into a series of 5 assumptions. This is similar to what you'd see in an econometrics textbook. This is helpful because later, we'll separately discuss how to assess each one.

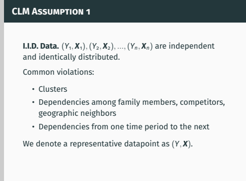
CLM Assumption 1

I.I.D. Data. $(Y_1, \boldsymbol{X}_1), (Y_2, \boldsymbol{X}_2), \ldots, (Y_n, \boldsymbol{X}_n)$ are independent and identically distributed.

Common violations:

- Clusters
- Dependencies among family members, competitors, geographic neighbors
- Dependencies from one time period to the next

We denote a representative datapoint as $(Y, \boldsymbol{X})$.



CLM Assumption 1

I.I.D. Data. $(Y_1, \boldsymbol{X}_1), (Y_2, \boldsymbol{X}_2), \ldots, (Y_n, \boldsymbol{X}_n)$ are independent and identically distributed.

Common violations:

- Clusters
- Dependencies among family members, competitors, geographic neighbors
- Dependencies from one time period to the next

We denote a representative datapoint as $(Y, \boldsymbol{X})$.

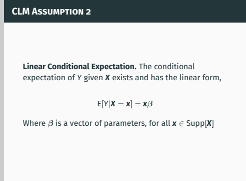
1. First, just like before, we need an assumption of I.I.D. iid tells us that each datapoint is 100% new information about the joint distribution.
2. The model is this: you have a joint distribution and you draw a datapoint. then you reset, totally clear memory and draw the next datapoint.
3. Real data generating processes very often don't look like that. Here are some of the most common violations
4. Some of these violations you can test for and model the dependency. for example, if you know what the clusters are, you can build them into a hierarchical model. but in general, you may now even know where all the dependencies are and have no way to model them.

CLM Assumption 2

Linear Conditional Expectation. The conditional expectation of Y given $\boldsymbol{X}$ exists and has the linear form,

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}\beta$$

Where β is a vector of parameters, for all $\boldsymbol{x} \in \text{Supp}[\boldsymbol{X}]$



CLM Assumption 2

Linear Conditional Expectation. The conditional expectation of Y given $\boldsymbol{X}$ exists and has the linear form,

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \boldsymbol{x}\beta$$

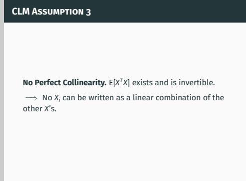
Where β is a vector of parameters, for all $\boldsymbol{x} \in \text{Supp}[\boldsymbol{X}]$

1. Next, we have linear conditional expectation.
2. This is quite strong. It's telling us that our joint distribution is essentially linear, although possible distributions around that line can still look very different.
3. This makes the regression problem much easier - we know we have a line, we just need to figure out which line.
4. commented out the zero-conditional mean version
5. Alex, can you remind students that $\boldsymbol{X} = [1, X_1, X_2, .., X_k]$

CLM Assumption 3

No Perfect Collinearity. $E[X^T X]$ exists and is invertible.

$\implies$ No X_i can be written as a linear combination of the other X's.



CLM Assumption 3

No Perfect Collinearity. $E[X^T X]$ exists and is invertible.

$\implies$ No X_i can be written as a linear combination of the other X's.

1. First, the mathematically precise definition...
2. What does this mean? For this to be invertible, that tells us that no x variable can be written as an exact linear combination of the other X's. It has to have some unique variation.
3. That makes sense because we know ols works on unique variation. If we combine this assumption with the previous one, we can prove that a unique BLP exists.

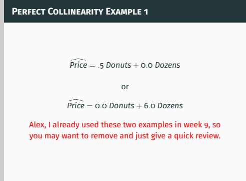
Perfect Collinearity Example 1

$$\widehat{Price} = .5\ Donuts + 0.0\ Dozens$$

or

$$\widehat{Price} = 0.0\ Donuts + 6.0\ Dozens$$

Alex, I already used these two examples in week 9, so you may want to remove and just give a quick review.



Perfect Collinearity Example 1

$\widehat{Price} = .5\ Donuts + 0.0\ Dozens$

or

$\widehat{Price} = 0.0\ Donuts + 6.0\ Dozens$

Alex, I already used these two examples in week 9, so you may want to remove and just give a quick review.

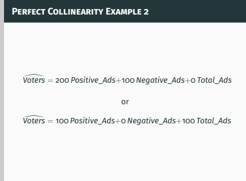
1. Here's an example to help understand why we need this assumption
2. You regress the price on both number of donuts and number of dozens of donuts.
3. it's 50 cents per donut, so you can write the price as 50 cents times number of donuts
4. But you can also write it as 6 dollars times number of dozens.
5. Both are equivalent for predicting the price - this problem doesn't affect prediction.
6. But you can't estimate coefficients, because they aren't uniquely defined.

Perfect Collinearity Example 2

$$\widehat{Voters} = 200\,Positive\_Ads + 100\,Negative\_Ads + 0\,Total\_Ads$$

or

$$\widehat{Voters} = 100\,Positive\_Ads + 0\,Negative\_Ads + 100\,Total\_Ads$$



Perfect Collinearity Example 2

$$\widehat{Voters} = 200\,Positive\_Ads + 100\,Negative\_Ads + 0\,Total\_Ads$$

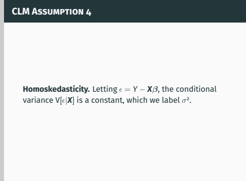
or

$$\widehat{Voters} = 100\,Positive\_Ads + 0\,Negative\_Ads + 100\,Total\_Ads$$

1. Here's another example, to show you that multicollinearity isn't always about pairs of variables, it might be more variables
2. Here you have a regression with number of positive ads, number of negative ads, and the total number of ads.
3. Once again, you can find multiple ways to write the same model.

CLM Assumption 4

Homoskedasticity. Letting $\epsilon = Y - \boldsymbol{X}\beta$, the conditional variance $\mathrm{V}[\epsilon|\boldsymbol{X}]$ is a constant, which we label σ^2.

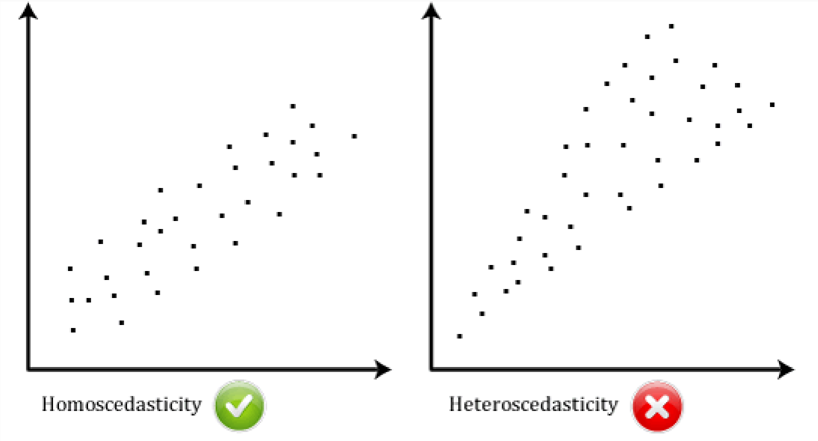


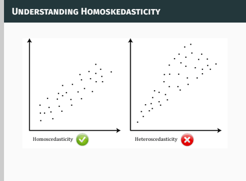
CLM Assumption 4

Homoskedasticity. Letting $\epsilon = Y - \boldsymbol{X}\beta$, the conditional variance $\mathrm{V}[\epsilon|\boldsymbol{X}]$ is a constant, which we label σ^2.

1. We already assumed we have a linear CEF, now we also need to know that the spread around that line is constant.
2. homo = same, skedasis = Greek for scattering

Understanding Homoskedasticity



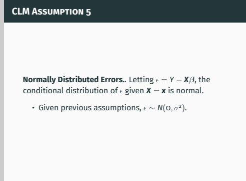

Understanding Homoskedasticity

1. borrowed image - make our own or fair use?
2. Here's a picture to show you what this assumption looks like.
3. To understand conditional variance, choose an X and then look at how wide the distribution is along the Y-axis.
4. On the left, we have an example of homoskedasticity - for any X, the conditional variance in Y looks the same.
5. On the right, we have heteroskedasticity, the conditional variance seems to increase towards the right.
6. This will complicate the behavior of our coefficients in a small sample, so we require homoskedasticity in the CLM.

CLM Assumption 5

Normally Distributed Errors.. Letting $\epsilon = Y - \boldsymbol{X}\beta$, the conditional distribution of ϵ given $\boldsymbol{X} = \boldsymbol{x}$ is normal.

- Given previous assumptions, $\epsilon \sim N(0, \sigma^2)$.



CLM Assumption 5

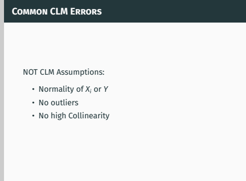
Normally Distributed Errors.. Letting $\epsilon = Y - \boldsymbol{X}\beta$, the conditional distribution of ϵ given $\boldsymbol{X} = \boldsymbol{x}$ is normal.

- Given previous assumptions, $\epsilon \sim N(0, \sigma^2)$.

Common CLM Errors

NOT CLM Assumptions:

- Normality of X_i or Y
- No outliers
- No high Collinearity


Common CLM Errors

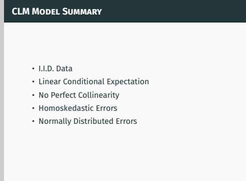
NOT CLM Assumptions:

- Normality of X_i or Y
- No outliers
- No high Collinearity

1. Finally, it's worth mentioning some common mistakes These are assumptions that are NOT part of the CLM.
2. First, there is no requirement that any variables are normal. Errors are normal. sometimes, if your variables are very skewed the error term will also be skewed, but that's not true 100% of the time.
3. Next, the CLM says nothing about outliers. Now, outliers may indicate that the error term is non-normal, but not necessarily. In any case, Never remove an outlier just because it's an outlier. You don't get to change the data just because it doesn't fit the model you want.
4. Finally, there is no assumption against multicollinearity, except for perfect multicollinearity. When

CLM Model Summary

- I.I.D. Data
- Linear Conditional Expectation
- No Perfect Collinearity
- Homoskedastic Errors
- Normally Distributed Errors



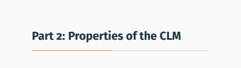
CLM Model Summary

- I.I.D. Data
- Linear Conditional Expectation
- No Perfect Collinearity
- Homoskedastic Errors
- Normally Distributed Errors

Part 2: Properties of the CLM

2020-07-22

└─Part 2: Properties of the CLM

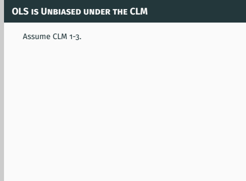

Part 2: Properties of the CLM

OLS is Unbiased under the CLM

OLS is Unbiased under the CLM

OLS is Unbiased under the CLM

Assume CLM 1-3.



OLS is Unbiased under the CLM

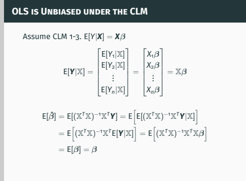
Assume CLM 1-3.

OLS is Unbiased under the CLM

Assume CLM 1-3. $\mathrm{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\beta$

$$\mathrm{E}[\boldsymbol{Y}|\mathbb{X}] = \begin{bmatrix} \mathrm{E}[Y_1|\mathbb{X}] \\ \mathrm{E}[Y_2|\mathbb{X}] \\ \vdots \\ \mathrm{E}[Y_n|\mathbb{X}] \end{bmatrix} = \begin{bmatrix} X_1\beta \\ X_2\beta \\ \vdots \\ X_n\beta \end{bmatrix} = \mathbb{X}\beta$$

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \mathrm{E}[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}] = \mathrm{E}\Big[\mathrm{E}[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}|\mathbb{X}]\Big] \\ &= \mathrm{E}\Big[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathrm{E}[\boldsymbol{Y}|\mathbb{X}]\Big] = \mathrm{E}\Big[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta\Big] \\ &= \mathrm{E}[\beta] = \beta \end{aligned}$$



OLS is Unbiased under the CLM

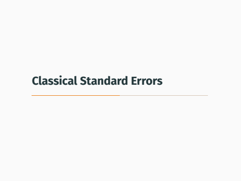
Assume CLM 1-3. $\mathrm{E}[Y|\boldsymbol{X}] = \boldsymbol{X}\beta$

$$\mathrm{E}[\boldsymbol{Y}|\mathbb{X}] = \begin{bmatrix} \mathrm{E}[Y_1|\mathbb{X}] \\ \mathrm{E}[Y_2|\mathbb{X}] \\ \vdots \\ \mathrm{E}[Y_n|\mathbb{X}] \end{bmatrix} = \begin{bmatrix} X_1\beta \\ X_2\beta \\ \vdots \\ X_n\beta \end{bmatrix} = \mathbb{X}\beta$$

$$\begin{aligned} \mathrm{E}[\hat{\beta}] &= \mathrm{E}[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}] = \mathrm{E}\Big[\mathrm{E}[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{Y}|\mathbb{X}]\Big] \\ &= \mathrm{E}\Big[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathrm{E}[\boldsymbol{Y}|\mathbb{X}]\Big] = \mathrm{E}\Big[(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta\Big] \\ &= \mathrm{E}[\beta] = \beta \end{aligned}$$

1. First, we need to write CLM 3 in matrix form. CLM 3 says that a single data point is expected to fall on the line, and we need to say that all the points fall on the line.
2. So you can see that the coefficients are unbiased, and for this we only needed CLM 1-3. The one strong assumption is linear conditional expectation.
3. This results is important, because it eliminates one way that our estimates can be wrong. There's no bias at any sample size. Next, we need a way to estimate variance in a small sample, and then we can start to understand our results.

Classical Standard Errors

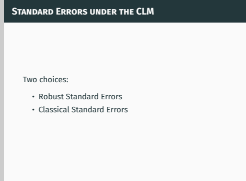


Classical Standard Errors

Standard Errors under the CLM

Two choices:

- Robust Standard Errors
- Classical Standard Errors



Standard Errors under the CLM

Two choices:

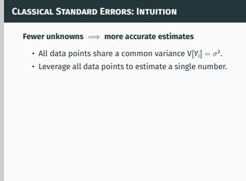
- Robust Standard Errors
- Classical Standard Errors

1. We already know that $\hat{\beta}$ is unbiased, but estimation is only the start of statistic. Now we turn to the issue of uncertainty - how can we quantify uncertainty in a small sample?
2. Let's start with the most fundamental metric of uncertainty: standard errors.
3. When working with the CLM, people use two different types of standard errors.
4. You've already seen robust standard errros, let's defines what classical standard errors are, and then talk about which of these you should use.

Classical Standard Errors: Intuition

Fewer unknowns $\implies$ more accurate estimates

- All data points share a common variance $\mathbb{V}[Y_i] = \sigma^2$.
- Leverage all data points to estimate a single number.



Classical Standard Errors: Intuition

Fewer unknowns $\implies$ more accurate estimates

- All data points share a common variance $\mathbb{V}[Y_i] = \sigma^2$.
- Leverage all data points to estimate a single number.

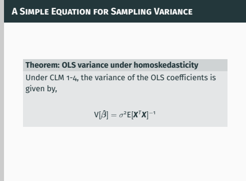
1. The basic idea is that under the CLM, we have fewer unknowns, so we should be able to use our data to get more accuracy
2. In particular, each datapoint has the same variance. Estimating n different variances would be hard.
3. since there's one variance, we can use every single datapoint to get more information about it
4. Here's a data point that's high - that's a signal that variance is high.
5. We can use a good estimate for error variance to compute the variance of the betas.

A Simple Equation for Sampling Variance

Theorem: OLS variance under homoskedasticity

Under CLM 1-4, the variance of the OLS coefficients is given by,

$$\mathrm{V}[\hat{\beta}] = \sigma^2 \mathrm{E}[\boldsymbol{X}^T \boldsymbol{X}]^{-1}$$



A Simple Equation for Sampling Variance

Theorem: OLS variance under homoskedasticity

Under CLM 1-4, the variance of the OLS coefficients is given by,

$$\mathrm{V}[\hat{\beta}] = \sigma^2 \mathrm{E}[\boldsymbol{X}^T \boldsymbol{X}]^{-1}$$

1. Let me remind you that the general equation for variance of betas is really complicated
2. But under homoskedasticity, we were able to simplify it a lot.
3. Here’s the theorem we saw earlier. we just multiply the error variance by this matrix.
4. How do we get an estimator for this. we just plug in sample analogues for variance and for the matrix

Classical Standard Errors

The classical variance estimator is

$$\widehat{\mathbb{V}}_C[\hat{\beta}] = \hat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$$

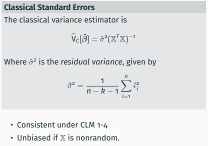
Where $\hat{\sigma}^2$ is the *residual variance*, given by

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

- Consistent under CLM 1-4
- Unbiased if $\mathbb{X}$ is nonrandom.

2020-07-22

└Classical Standard Errors



Classical Standard Errors

The classical variance estimator is

$$\widehat{\mathbb{V}}_C[\hat{\beta}] = \hat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$$

Where $\hat{\sigma}^2$ is the *residual variance*, given by

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{\epsilon}_i^2$$

- Consistent under CLM 1-4
- Unbiased if $\mathbb{X}$ is nonrandom.

1. This is what we call classical standard errors.
2. For us, these are consistent, but textbooks that use stronger assumptions list them as unbiased.

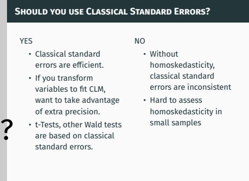
Should you use Classical Standard Errors?

YES

- Classical standard errors are efficient.
- If you transform variables to fit CLM, want to take advantage of extra precision.
- t-Tests, other Wald tests are based on classical standard errors.

NO

- Without homoskedasticity, classical standard errors are inconsistent
- Hard to assess homoskedasticity in small samples



Should you use Classical Standard Errors?

YES

- Classical standard errors are efficient.
- If you transform variables to fit CLM, want to take advantage of extra precision.
- t-Tests, other Wald tests are based on classical standard errors.

NO

- Without homoskedasticity, classical standard errors are inconsistent
- Hard to assess homoskedasticity in small samples

1. In simulations that we ran, classical standard errors had about half the variance of robust standard errors.
2. For these reasons, there are statisticians that recommend always using robust standard errors. It’s not a bad policy, but you can look at the tradeoff and make your own decision. Just make sure the evidence for homoskedasticity is strong before you use classical standard errors.

Sampling Distributions under the CLM

Sampling Distributions under the CLM

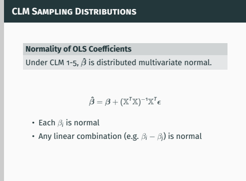
CLM Sampling Distributions

Normality of OLS Coefficients

Under CLM 1-5, $\hat{\beta}$ is distributed multivariate normal.

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}$$

- Each β_i is normal
- Any linear combination (e.g. $\beta_i - \beta_j$) is normal


CLM Sampling Distributions
Normality of OLS Coefficients
Under CLM 1-5, $\hat{\beta}$ is distributed multivariate normal.
$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\boldsymbol{\epsilon}$$
- Each β_i is normal
- Any linear combination (e.g. $\beta_i - \beta_j$) is normal

1. To do more precise inference, including confidence intervals and tests, we have to know what the sampling distribution of our beta hats is.
2. We need the full CLM for this. The result is that under CLM 1-5, our beta hats are normal.
3. Here's how you know. remember this equation for beta hat...
4. Draw arrow for epsilon and label normal.
5. ϵ is normal, and we pass it through a linear transformation. so the result is also normal.

Testing Coefficients under the CLM

Theorem: t-statistics for OLS Coefficients

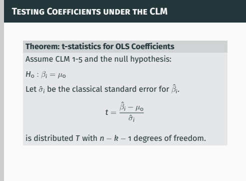
Assume CLM 1-5 and the null hypothesis:

$H_0 : \beta_i = \mu_0$

Let $\hat{\sigma}_i$ be the classical standard error for $\hat{\beta}_i$.

$$t = \frac{\hat{\beta}_i - \mu_0}{\hat{\sigma}_i}$$

is distributed T with $n - k - 1$ degrees of freedom.



Testing Coefficients under the CLM

Theorem: t-statistics for OLS Coefficients

Assume CLM 1-5 and the null hypothesis:

$H_0 : \beta_i = \mu_0$

Let $\hat{\sigma}_i$ be the classical standard error for $\hat{\beta}_i$.

$$t = \frac{\hat{\beta}_i - \mu_0}{\hat{\sigma}_i}$$

is distributed T with $n - k - 1$ degrees of freedom.

1. Now that we know that our sampling distribution is normal, we can derive a t-statistic
2. The setup is very similar to a classic t-test for one variable.
3. We have a null hypothesis, which is usually that our β_i is zero.
4. We standardize, by subtracting the hypothesized mean, and dividing by the standard error.
5. Since we're estimating the standard error, the statistic doesn't follow a z-distribution, it follows a t-distribution.
6. By the way, the degrees of freedom is n-k, we lose a degree of freedom for every beta we estimate.
7. So this is why we always run t-tests instead of z-tests in a

Take-Aways

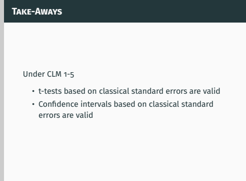
Under CLM 1-5

- t-tests based on classical standard errors are valid
- Confidence intervals based on classical standard errors are valid

2020-07-22

└─Sampling Distributions under the CLM

└─Take-Aways



Take-Aways

Under CLM 1-5

- t-tests based on classical standard errors are valid
- Confidence intervals based on classical standard errors are valid

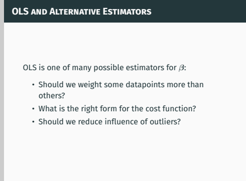
Efficiency Theorems for OLS

Efficiency Theorems for OLS

OLS and Alternative Estimators

OLS is one of many possible estimators for β:

- Should we weight some datapoints more than others?
- What is the right form for the cost function?
- Should we reduce influence of outliers?



OLS and Alternative Estimators

OLS is one of many possible estimators for β:

- Should we weight some datapoints more than others?
- What is the right form for the cost function?
- Should we reduce influence of outliers?

1. I want to talk about efficiency, but first, let me remind you that there are many possible estimators out there.
2. OLS is simple and elegant, but we should remember that it’s just one choice out of many
3. If you take a step back and think about how you would design an algorithm to estimate the true betas, here are a few things you might consider.
4. Should we weight some datapoints more than others - maybe those that appear least noisy...
5. If you think about these choices, you could write different algorithms to take the place of ols - is that a good idea?

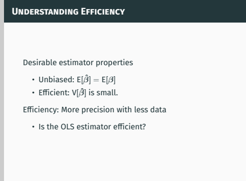
Understanding Efficiency

Desirable estimator properties

- Unbiased: $E[\hat{\beta}] = E[\beta]$
- Efficient: $V[\hat{\beta}]$ is small.

Efficiency: More precision with less data

- Is the OLS estimator efficient?



Understanding Efficiency

Desirable estimator properties

- Unbiased: $E[\hat{\beta}] = E[\beta]$
- Efficient: $V[\hat{\beta}]$ is small.

Efficiency: More precision with less data

- Is the OLS estimator efficient?

1. When you're comparing estimators, it's good to go back and review the properties that we want estimators to have. Here's two...
2. One is being unbiased, and we know ols is unbiased
3. Another is what Statisticians call efficiency. An estimator is efficient if it has low variance.
4. That really means efficient with data, getting more precision with less data.
5. your standard errors will be small, you'll have good precision, your hypothesis tests will be powerful...
6. Is the OLS estimator efficient? There are some famous theorems about this. I'm going to give you a very brief overview of two

The Gauss-Markov Theorem

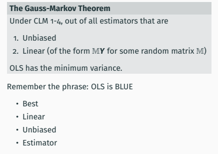
Under CLM 1-4, out of all estimators that are

1. Unbiased
2. Linear (of the form $\mathbb{M}\boldsymbol{Y}$ for some random matrix $\mathbb{M}$)

OLS has the minimum variance.

Remember the phrase: OLS is BLUE

- Best
- Linear
- Unbiased
- Estimator



The Gauss-Markov Theorem

Under CLM 1-4, out of all estimators that are

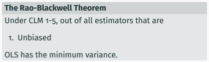
1. Unbiased
2. Linear (of the form $\mathbb{M}\boldsymbol{Y}$ for some random matrix $\mathbb{M}$)

OLS has the minimum variance.

Remember the phrase: OLS is BLUE

- Best
- Linear
- Unbiased
- Estimator

1. First, we have the famous Gauss-Markov theorem. These guys looked at all estimators that are unbiased and also linear, meaning they can be written as some matrix times Y. Our of all those estimators, ols has minimum variance.
2. You need the first 4 CLM assumptions for this to be true, including homoskedasticity. Without homoskedasticity, a different algorithm, called weighted least squares is more efficient.



The Rao-Blackwell Theorem

Under CLM 1-5, out of all estimators that are

1. Unbiased

OLS has the minimum variance.

The Rao-Blackwell Theorem

Under CLM 1-5, out of all estimators that are

1. Unbiased

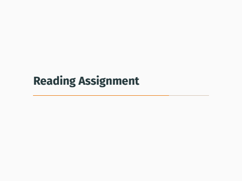
OLS has the minimum variance.

1. There's also the Rao-Blackwell theorem, that says that under all 5 CLM assumptions, if you look at unbiased estimators, ols has the minimum variance.
2. You can take these as evidence that ols is efficient. As a data scientist, you probably don't need more reasons to use ols, you just use it. But this is a peek at how statisticians think about estimators that can help you as you learn more statistics.
3. Talk about how these theorems are a little silly. because they apply only to a very restrictive setting.

Reading Assignment

2020-07-22

└─Reading Assignment



Reading Assignment

Reading Assignment

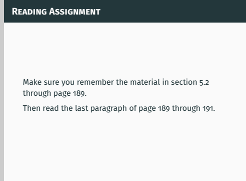
Make sure you remember the material in section 5.2 through page 189.

Then read the last paragraph of page 189 through 191.

2020-07-22

└─Reading Assignment

└─Reading Assignment

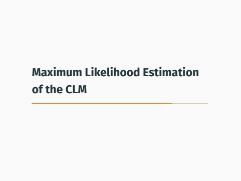


Reading Assignment

Make sure you remember the material in section 5.2 through page 189.

Then read the last paragraph of page 189 through 191.

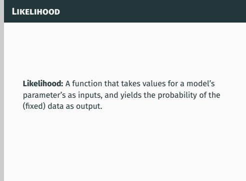
Maximum Likelihood Estimation of the CLM



Maximum Likelihood Estimation of the CLM

Likelihood

Likelihood: A function that takes values for a model's parameter's as inputs, and yields the probability of the (fixed) data as output.



Likelihood

Likelihood: A function that takes values for a model's parameter's as inputs, and yields the probability of the (fixed) data as output.

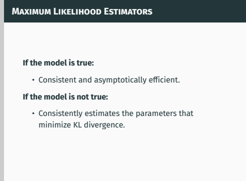
Maximum Likelihood Estimators

If the model is true:

- Consistent and asymptotically efficient.

If the model is not true:

- Consistently estimates the parameters that minimize KL divergence.



Maximum Likelihood Estimators

If the model is true:

- Consistent and asymptotically efficient.

If the model is not true:

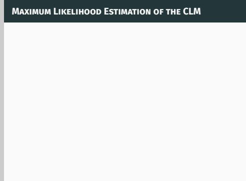
- Consistently estimates the parameters that minimize KL divergence.

Maximum Likelihood Estimation of the CLM

2020-07-22

└─Maximum Likelihood Estimation of the CLM

└─Maximum Likelihood Estimation of the CLM



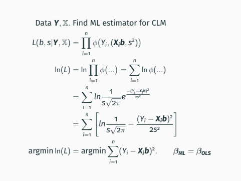
Maximum Likelihood Estimation of the CLM

Data $\boldsymbol{Y}, \mathbb{X}$. Find ML estimator for CLM

$$L(b, s|\boldsymbol{Y}, \mathbb{X}) = \prod_{i=1}^{n} \phi(Y_i, (\boldsymbol{X_i b}, s^2))$$

$$\begin{aligned}
\ln(L) &= \ln \prod_{i=1}^{n} \phi(\ldots) = \sum_{i=1}^{n} \ln \phi(\ldots) \\
&= \sum_{i=1}^{n} ln \frac{1}{s\sqrt{2\pi}} e^{\frac{-(Y_i - \boldsymbol{X_i b})^2}{2s^2}} \\
&= \sum_{i=1}^{n} \left[ln \frac{1}{s\sqrt{2\pi}} - \frac{(Y_i - \boldsymbol{X_i b})^2}{2s^2} \right]
\end{aligned}$$

$$\text{argmin} \ln(L) = \text{argmin} \sum_{i=1}^{n} (Y_i - \boldsymbol{X_i b})^2. \qquad \beta_{ML} = \beta_{OLS}$$



Data $\boldsymbol{Y}, \mathbb{X}$. Find ML estimator for CLM

$L(b, s|\boldsymbol{Y}, \mathbb{X}) = \prod_{i=1}^{n} \phi(Y_i, (\boldsymbol{X_i b}, s^2))$

$\ln(L) = \ln \prod_{i=1}^{n} \phi(\ldots) = \sum_{i=1}^{n} \ln \phi(\ldots)$

$= \sum_{i=1}^{n} ln \frac{1}{s\sqrt{2\pi}} e^{\frac{-(Y_i - \boldsymbol{X_i b})^2}{2s^2}}$

$= \sum_{i=1}^{n} \left[ln \frac{1}{s\sqrt{2\pi}} - \frac{(Y_i - \boldsymbol{X_i b})^2}{2s^2} \right]$

$\text{argmin} \ln(L) = \text{argmin} \sum_{i=1}^{n} (Y_i - \boldsymbol{X_i b})^2. \qquad \beta_{ML} = \beta_{OLS}$

1. Conclusion: We get exactly the ols estimator. In this case, the ML estimator and the ols estimator are the same. This is another way that you can motivate the ols algorithm.
2. By the way, remember one cool fact about ML estimators. even if the model is not true, ML will consistently find the model that's as close as possible to the true distribution, where close is defined as KL-divergence. So that means that even if the CLM does hold, we're estimating betas that give as good as approximation as possible.
3. We usually prefer to use as few assumptions as possible, so this isn't our favorite way to think about ols. But we want you to see it from this perspective, so you can connect it to the models you'll see in ML or in your

Can We Believe the CLM?

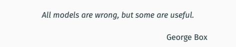
└─Can We Believe the CLM?

Can We Believe the CLM?

All models are wrong, but some are useful.

George Box

Can We Believe the CLM?



All models are wrong, but some are useful.

George Box

1. If we assume the CLM, then we get a lot of great statistical guarantees - guarantees that let us work with small samples.
2. But assumptions are not true because we want them to be true.
3. the CLM in particular is a parametric model - a very strict set of assumptions. Can we really believe it?
4. Clearly the textbook authors are skeptics. The repeatedly say that you can't take the CLM literally.
5. Alex and I think they make a good point. Of course you should be skeptical of any model. No model is actually TRUE - it's a model. What's important is to examine the model and see if it reminds us of the real world, or if we

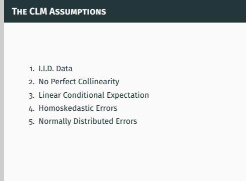
The CLM Assumptions

1. I.I.D. Data
2. No Perfect Collinearity
3. Linear Conditional Expectation
4. Homoskedastic Errors
5. Normally Distributed Errors

2020-07-22

└Can We Believe the CLM?

└The CLM Assumptions



The CLM Assumptions

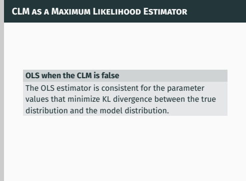
1. I.I.D. Data
2. No Perfect Collinearity
3. Linear Conditional Expectation
4. Homoskedastic Errors
5. Normally Distributed Errors

1. So we think that you should indeed be skeptical of the CLM. But it's important to qualify that statement. First, we shouldn't lump the entire CLM in together
2. We provided 5 assumptions. and as you go from top to bottom, adding one at a time, the model gets more restrictive
3. Remember that some guarantees require all 5 assumptions, and some don't. unbiased coefficients only require 1-3. and you can usually transform your data so that the linear conditional expectation assumption looks plausible.
4. t-Tests require all 5 assumptions, and we think that you should definitely be more skeptical about whether

CLM as a Maximum Likelihood Estimator

OLS when the CLM is false

The OLS estimator is consistent for the parameter values that minimize KL divergence between the true distribution and the model distribution.



CLM as a Maximum Likelihood Estimator

OLS when the CLM is false

The OLS estimator is consistent for the parameter values that minimize KL divergence between the true distribution and the model distribution.

1. Another important point that the authors make, is that even if the CLM is not true, it’s a max likelihood estimator.
2. So that means that it consistently estimates a distribution that’s as close to the true distribution as possible.
3. So even if the CLM is false, OLS doesn’t totally give up. It still tries to get as close as possible. Of course, we need a somewhat large n for this to be meaningful.

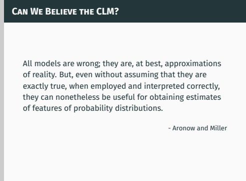
Can We Believe the CLM?

> All models are wrong; they are, at best, approximations of reality. But, even without assuming that they are exactly true, when employed and interpreted correctly, they can nonetheless be useful for obtaining estimates of features of probability distributions.
>
> - Aronow and Miller

2020-07-22

└─Can We Believe the CLM?

└─Can We Believe the CLM?



Can We Believe the CLM?

All models are wrong; they are, at best, approximations of reality. But, even without assuming that they are exactly true, when employed and interpreted correctly, they can nonetheless be useful for obtaining estimates of features of probability distributions.

- Aronow and Miller

Assessing the CLM assumptions - Software Demo

Assessing the CLM assumptions - Software Demo

This one will be a screenshare using R Studio

└─Assessing the CLM assumptions - Software Demo

This one will be a screenshare using R Studio