Boston University Questrom School of Business

MF793 – Fall 2021

Eric Jacquier

OLS: When things go wrong

A. Quick Return on Correlated Regressors

B. Omitted X variables

C. RHS variable measured with error

D. Using the residuals as diagnostic of the (unknown) noise properties

E. Difference between residual and forecast error, the leave-one-out residual

F. Errors are not i.i.d: Heteroskedastic and / or correlated errors,

Readings: Hansen Ch. 2, 3, 4

Greene: Ch. 4.7

A. Quick return on correlated regressors

Recall $\mathbf{V(\hat{\beta}_{OLS}) = \sigma^2 (X'X)^{-1}}$

Highly correlated X variables increase the variance of the OLS. .. Which coefficient?

Is there some intuition in $(X'X)^{-1}$ diagonal elements?

Yes, we can show (no proof) that:

$$V(\hat{\beta}_{OLS,k}) \equiv \sigma^2 \{(X'X)^{-1}\}_{kk} = \frac{\sigma^2}{(1-R_k^2)\sum_1^T (x_{k,i}-\overline{x_k})^2}$$

Where:

R_k^2 is the R2 of the regression of x_k on all the other x's

$\frac{1}{(1-R_k^2)}$ is called the **variance inflation factor** of the estimate of β_k.

B. Missing Variable, we regress Y on X_1, forget a variable X_2

B1 Effect on the coefficient estimate

Say the correct model is: $Y = X_1\beta_1 + X_2\beta_2 + \varepsilon^*,\ E(\varepsilon^* \mid X_1, X_2) = 0$ [1]

But we think ***incorrectly*** that: $Y = X_1\beta + \varepsilon,$ [2]

Our estimate from the wrong model is: $b = (X_1'X_1)^{-1} X_1'Y$

When is b an unbiased estimate of β_1?

$b = (X_1'X_1)^{-1} X_1'$ **Y** $= (X_1'X_1)^{-1} X_1' (X_1\beta_1 + X_2\beta_2 + \varepsilon)$ <- Y by true model

$= \beta_1 + (X_1'X_1)^{-1} X_1'X_2\beta_2 + (X_1'X_1)^{-1} X_1'\varepsilon$

$= \beta_1 + (X_1'X_1)^{-1} X_1'\varepsilon + (X_1'X_1)^{-1} X_1'X_2\beta_2$

$E(\boldsymbol{b}) = \beta_1 + + \mathbf{(X_1'X_1)^{-1} X_1'X_2}\beta_2$

$\mathbf{(X_1'X_1)^{-1} X_1'X_2}$: The matrix of the regression coefficients X_2 on X_1

=> **The estimate *b* is biased iff X_1 is uncorrelated with X_2**
That is if the regression coefficient of X_2 on X_1 is zero.

B2 Risk of too many vs Risk of not enough variables

- Omitting variables:
  - $\hat{\beta}$ for the included variables can be a biased estimate of β_1
    We can also show that variance of the estimate is smaller (than if using correct model) ! no proof

  - Effect of omitting X2 on a forecast $\hat{y}$ is less obvious
    Biased coefficient estimate in effect compensates (not totally) for missing variable

- Do we run the regression to test a hypothesis on β_1 or to forecast y?

- Contrast with including too many *"not so useful"* variables!
  If the included variables are (highly) correlated
  - variance of the coefficient estimator increases
  - variance of the forecast error increase
  - No bias
  - The inference is correct, both for estimation and forecast

- Boils down to modeling strategy: risk of too many variables vs risk of missing essential variables.
  - Old days: Bottom up modeling favored, start with few X variables, build up the model
  - Less old days: Top down preferred, start from many variables, reduce the model.
  - Either way is deeply flawed.

    Both introduce severe biases known as "**pretest biases**" or "**data mining biases**"

- What are data mining biases:
  - Say your final model is: Y on X_1
  - Your final model *won* after many trials of Y on X_1 .. , and on X_2, X_3, .. which were eliminated.
  - Standard t-test for $\hat{\beta}_1$ in your final model **overestimates** its significance.

Think: Assume you start with 100 candidate X variables, trying to forecast Y

- What to do ?
  - **No mechanical variable selection technique is reliable …**

    … without a solid dose of common sense and subject matter knowledge.
  - Know your subject, don't use mechanical model selection techniques

  - If you must do model selection, avoid criteria that lead to large models.
    R^2 is useless for model selection, adjusted R^2 is hardly better.
    Prefer AIC or SIC criteria (will see in Time series)

  - If you did any model selection, you can not use the usual standard error and t-statistic for your estimates and forecasts.

    because the **data-mining** you conducted modified the *"nominal"* distribution of your test

  - Even better: model combination rather than model selection, odds ratios.

    Why "throw away" models?

    Does a portfolio manager throw away the stocks with higher variance?

C Variables are measured with errors

C1 Dependent variable Y measured with error

- Theoretical Model: $Y^* = X\beta + \varepsilon$

  We don't know Y*, we **measure** it: $Y = Y^* + \nu$ ν: random measurement error

  So the model for the data Y is: $Y = X\beta + \varepsilon + \nu$ [1]

- If the measurement error is additive, unrelated to X, and homoskedastic,

  the only consequence is a higher variance of noise when we use Y

  => lower fit.

  But the model's assumption are maintained, the inference is correct.

C2 Independent variable X measured with error

- Theoretical Model: $Y = \alpha + X^*\beta + \varepsilon$

  We don't know X*, we measure it: $X = X^* + \nu$ $\nu \sim (0, \sigma_\nu)$

  So the model for the data X,Y is: $Y = \alpha + X\beta + (\varepsilon - \beta\nu)$ [2]

- **$Cov(\varepsilon - \beta\nu, X)$** $= Cov(\varepsilon, X) - \beta\, Cov(\nu,X) = 0 - \beta\, Cov(X^*+\nu,\nu) =$ **$-\beta\, Var(\nu) \neq 0$**

- Even for the simplest assumption on the measurement error: $Cov(\nu, X^*) = Cov(\nu, \varepsilon) = 0$,

  the errors in the model we run ([2]) are correlated with the RHS variable we use, X.

  The OLS estimate is biased.

- We can also show that the bias of the OLS estimate of β no proof
  - is generally toward zero,
  - is larger the larger the size of measurement error σ_ν .

When the RHS variables are estimated with error,

the OLS beta estimator is generally biased toward zero,

the bias is larger the larger σ_ν

D Using the residuals to conduct model diagnostic

- We use the residuals *e* to diagnose
  1) patterns of dependencies between X and ε
  2) heteroskedasticity of the noise
  3) patterns of dependencies among **ε**
- Let's understand the distribution of these residuals
- Covariance matrix of the residuals:

Recall $e = M\varepsilon$ where $M = I - X(X'X)^{-1}X' = I - P$ M is symmetric

$$\mathbf{V(e)} = E(ee') = E(M\,\varepsilon\,\varepsilon'\,M') = M\,E(\,\varepsilon\,\varepsilon')\,M = \mathbf{\sigma^2\,M \neq \sigma^2\,I} \qquad \mathbf{[1]}$$

When the true noise is i.i.d., the residuals

are **not uncorrelated** with one another and **not homoskedastic!**

Variance of one residual:

$$\mathbf{V(e_i)} = \sigma^2\,m_{ii} = \mathbf{\sigma^2\,(\,1 - p_{ii})} \qquad \mathbf{[1']}$$

- Result: Standardized residuals **can not be computed as** $\frac{e_i}{s}$ even if s^2 is an unbiased estimate of σ^2.

- Standardized residuals **must be computed as**

$$\frac{e_i}{s\sqrt{1-p_{ii}}} \qquad \textbf{[2]}$$

  Residuals as in [2] are sometimes called **studentized** residuals.

  Be careful when reading / using different packages !

- So, if we call [2] a standardized residual, what is a studentized residuals ?

  It is a (third!) standardization where s in [2] is replaced by s_i, an estimate of σ which excludes the residual i. … That is we leave observation i out to compute s_i.

  There is no common standard used by everybody to refer to the standardized (or studentized?) residuals!

- Never mind the jargon, you should know [1'] and use [2]. How is it done in R:

  influence(model)$hat

  gives the vector of diagonals values of P, the **Hat** matrix

  plot(model)

  gives diagnostic plots based on standardized residuals computed with [2]

- It's a good time to discuss *"leave one out"* diagnostics.

E. Leave-one-out and all that sort of things. (Hansen 3.16 for details)

- What is the difference between a residual and a forecast error?
  Residual: $e_i = y_i - x_i\hat{\beta} = y_i - \hat{y}_i$ Forecasting error: $e_f = y_f - x_f\hat{\beta} = y_f - \widetilde{y_f}$
  Observation *i* was used to estimate β **Observation *f* was not !**

- Say we use T observations to estimate β, and we have T_2 additional "***out-of-sample***" observations.
  - We can compute an **Out-Of-Sample R^2** with the T_2 observations which **did not** influence $\hat{\beta}$
  - OOS R^2 much lower than the in-sample R^2 indicates model instability … or data mining. ☠

  Save data for OOS forecasting, as diagnostic of stability / data mining on your model.

- The **Leave-one-out** technique.
  - For each (x'_i, y_i), estimate β by **excluding** observation I, -> $\widehat{\beta_{-\iota}}$, $\widetilde{y}_\iota$, $\widetilde{e}_\iota$ ***Leave-one-out***
  - Can show: $\hat{y}_\iota - \widetilde{y}_\iota = p_{ii}\,\widetilde{e}_\iota$. Diagonal elements of the P matrix diagnose influence point.
  - Compute R^2 with the leave-one-out fitted values and residuals.
  - The ***leave-one-out R^2*** is a better model selection criterion than R^2 and $\bar{R}^2$

- **Cross-validation**
  Large models vastly **over**fit: They achieve very high in-sample R^2 which collapse out of sample.
  Say: T observations, Leave K out, estimate on T-K
  Only 1 way to leave-one-out, MANY way to leave k out
  Cross validation is a generalization of "leave one out"

F Heteroskedasticity, Correlated errors, GLS or adjusted OLS Standard errors?

F1. Complication: $E(\varepsilon\varepsilon') \neq \sigma^2 I$ $E(\varepsilon\varepsilon') = \sigma^2 \Omega = \Sigma$

- We write $\sigma^2 \Omega$ so that the general model nests the iid model.

  But of course in many situations, σ is not identified uniquely. It's only a convention.

- Heteroskedasticity: $E(\varepsilon\varepsilon') = D \neq \sigma^2 I$, $E(\varepsilon_i^2) \neq E(\varepsilon_j^2)$

$$\sigma^2 \mathbf{\Omega} = \sigma^2 \begin{bmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \omega_n \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ & & \vdots & \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}.$$

Examples:

- Error variance proportional to one of the regressors: $\sigma_i = \sigma\sqrt{x_i}$
- Error variance ... varies with time or previous shock: $\sigma_t = \alpha + \delta r_{t-1}^2$ ARCH model

- Correlated errors $E(\varepsilon_I \varepsilon_j) \neq 0$ for $i \neq j$, $E(\varepsilon\varepsilon') = \sigma^2 \Omega$

  For homoskedastic, correlated errors, σ^2 is identified. It the constant error variance and the Ω matrix is a correlation matrix.

- Examples:
  - Cross-sectional regression where observations indicate for example industry / region

    Group effects: $i \neq j$ in same industry/region, $E(\varepsilon_i \varepsilon_j) \neq 0$

    $i \neq j$ in different industries/regions $E(\varepsilon_i \varepsilon_j) = 0$

  - Time Series regression

    Autocorrelation of order 1: $\varepsilon_t = \rho\, \varepsilon_{t-1} + \nu_t$ AR(1)

$$\sigma^2 \mathbf{\Omega} = \sigma^2 \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \cdots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \cdots & 1 \end{bmatrix}$$

where $\rho_K = \text{Corr}(\varepsilon_t, \varepsilon_{t-k})$. We expect ρ to be higher if k is smaller (close observations)

We will see that for the AR(1) model $\rho_K = \rho^K$

F2. General Solution: Generalized Least Squares (GLS)

- Matrix Result: $\forall\ \Omega, \exists\ P\ /\quad \Omega = P'\ D\ P$ [1]

  $P'P = P\ P' = I$ and D is (the) diagonal (matrix of *Eigen values*).

- Pre-multiply Ω by $P^* = D^{-0.5}P$

  $D^{-0.5}\ P\ \ \Omega\ \ P'\ \ D^{-0.5} = D^{-0.5}\ P\ \ P'DP\ \ P'\ \ D^{-0.5} = \ D^{-0.5}\ D\ \ D^{-0.5} = I_T$

  $P^*\ \ \Omega\ \ P^{*\prime} = I_T$ Keep this ready near the stove.

- Pre-multiply our regression: $P^*\ Y = P^*\ X\ \beta + P^*\ \varepsilon$ $\qquad E(\varepsilon\varepsilon') = \sigma^2\Omega$

  $Y^* = X^*\ \beta + \varepsilon^*$ [2]

- **$E(\varepsilon^*\varepsilon^{*\prime}) = E\ (P^*\ \varepsilon\ \varepsilon'\ P^{*\prime}) = P^*\ \sigma^2\ \Omega\ \ P^{*\prime} = \sigma^2\ I$**

  => OLS applies to [2], it is BLUE for [2]

- GLS is just OLS applied to the transformed system [2]

  $\hat{\beta}_{GLS} = (X^{*\prime}\ X^*)^{-1}\ X^{*\prime}\ Y^* = (X'\Omega^{-1}X)^{-1}\ X'\Omega^{-1}\ Y$ [3]

  $V(\hat{\beta}_{GLS}) = \sigma^2\ (X^{*\prime}X^*)^{-1} = \sigma^2\ (X'\Omega^{-1}X)^{-1}$ [4]

- Are we done? No! Not even started!
  We need an estimate of Ω to plug into [3][4]: $\widehat{\Omega}$ to get a *Feasible GLS*

- Ω is a TxT matrix.

  Under i.i.d. errors, we reduced it to one parameter, $\sigma^2 I_T$.

  We now have T(T+1)/2 parameters and only T observations

- **Can't** estimate T(T+1)/2 parameters with T observations !

  Need to specify a model of correlation / heteroskedasticity to reduce the number of parameters

  Examples: autocorrelated residuals with AR(1) — One parameter ρ describes entire matrix
  Weighted Least Squares, $\sigma_i = \sigma x_i$ — Observable describes variance
  Noise variance follows a GARCH process

- **Problem:** GLS is very sensitive to assumptions on the covariance matrix

  Wrong $\widehat{\Omega}$ means wrong $\hat{\beta}_{GLS}$: GLS can then be biased and inefficient

  The GLS variance covariance matrix is then incorrect: $V(\hat{\beta}_{GLS}) \neq \sigma^2 (X'\Omega^{-1}X)^{-1}$

GLS is a nice idea but you need to be confident that your model is well specified.

F3. Remain with OLS, get a good estimate of the OLS standard error. HAC standard errors

What is wrong with OLS if $E(\varepsilon\varepsilon') = \sigma^2\Omega$?

- OLS is likely inefficient.
- But we are not sure of what model to use for GLS.
- The variance of $\hat{\beta}_{OLS}$ is **not** $\sigma^2(X'X)^{-1}$.

That formula is wrong, bye-bye confidence intervals for estimates and forecasts

Possible view:

- GLS is tricky: if we pick the wrong Ω, the remedy may be worse than the disease.
- Better a somewhat inefficient estimator (OLS) with properly computed standard errors than a mis-specified estimator with incorrectly computed everything.

Counter View:

At least for Time-Series models, one can easily devise a GLS based AR model + GARCH model for the errors. It would likely dominate OLS.
This is true: We will do this in MF840

- For now, we introduce the **robust standard errors** for:

Unspecified heteroskedasticity Hal White (1983)
Unspecified autocorrelation Newey West (1987).

They allow us to keep using OLS while correcting the variance formula

- For heteroskedasticity: $E(\varepsilon\varepsilon') = D$. We don't know D!

$V(\hat{\beta}_{OLS}) = E[(X'X)^{-1} X' \varepsilon \varepsilon' X (X'X)^{-1}] = \sigma^2 (X'X)^{-1} X' D X (X'X)^{-1}$

Ideal estimator if we knew the true noises: $X'\varepsilon\varepsilon'X = \sum_i x_i x_i' \varepsilon_i^2$

Hal White (1983):

We can estimate $E(X' \varepsilon \varepsilon' X)$ consistently even if we can't estimate D consistently

We replace the true ε_i's with the residual e_i's or better, rescaled e_i's.

Of course, it does not estimate D consistently, $\{diag(e^2_i)\}$, one observation per parameter!

But it estimates $X' D X$ and hence $V(\hat{\beta}_{OLS})$ which are KxK

- Autocorrelated Errors: Hansen-Hodrick and Newey and West, propose similar strategies. They use ad-hoc estimates of autocorrelation until fairly high lags

Hansen-Hodrick worked in a context of overlapping observations where they **knew** how many lags of autocorrelations were in the data.

In general, it can create non-positive covariance matrices. Newey-West corrects this problem and is preferred.

- These are called **HAC standard errors**: **H**eteroskedasticity – **A**utocorrelation – **C**onsistent
  OLS may not be the "best" estimator but you can specify its uncertainty properly

- **In R:** command coeftest, needs packages sandwich and lmtest.