MA Quant II: Final Review[a]

Felipe Balcazar

NYU

May 3, 2021

[a] Thanks to Ali Ahmed

What are we doing?

- Observe the world.
- Intuit a relationship b/t Y and $X = (X_1, X_2, ..., X_p)$.

$$Y = f(X) + \epsilon$$

- Y = *systematic component* (S) + *random error term* (E)
- S: systematic information that X provides about Y.
- E: random error term with mean 0 and independent of X.
- Task: estimate f using random samples from the population.

How do we estimate f?

There are many methods. All boil down to:

1. Sample data (preferably random).
2. Use sampled data to estimate f.

Two approaches:

1. Parametric:
   - Make assumptions on the functional form of f.
   - Estimate parameters using sampled data.
2. Non-Parametric:
   - Doesn't rely on assumptions about f.
   - Often times more accurate but less interpretable.

Recall Gauss Markov assumptions

1. $Y_i = \beta X_i + \varepsilon_i$
2. $E[(\varepsilon_i|X_i)] = 0$
3. $Cov(\varepsilon_i, \varepsilon_j) = 0 \forall i \neq j$
4. $Var(\varepsilon_i|X_i) = \sigma^2 \forall i$
5. $Cov(\varepsilon_i, X_i) = 0 \forall i$
6. $Cov(X_i, X_j) \neq 1 \forall i \neq j$ and $X < N$

English:

1. For any value of X the disturbances average out to 0
2. The disturbance term is independent across observations
3. The variance of the disturbance term is the same for all i
4. The disturbances are exogenous
5. The regression model is properly specified
6. No exact linear relationship b/t X's and more obs than X's

Normality assumption

To make inferences on the coefficient estimates we assume:

$$\epsilon_i \sim N(0, \sigma^2)$$

From this assumption it follows that:

$$\hat{\beta}_p \sim N(\beta_p, \sigma_{\hat{\beta}_p})$$

This assumption allows us to perform hypotheses test.

Assumption violations

We have seen four:

1. Model miss-specification
   - $Y_i \neq \beta X_i + \varepsilon_i$
   - $Cov(\varepsilon_i, X_i) \neq 0$ for some X_i
2. Heteroskedasticity
   - $Var(\varepsilon_i | X_i) \neq \sigma^2$ for some X_i
3. Measurement error
   - Potentially: $Cov(\varepsilon_i, X_i) \neq 0$ for some X_i
4. Multicollinearity
   - $Cov(X_i, X_j) \approx 1$ for some $X_i \neq X_j$

Misspecification: types of misspecifications

1. Omitted variables
   - True model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \varepsilon_i$
   - Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
   - If $Cov(X_1, X_2) \neq 0 \rightarrow$ biased coefficients and std. errors.
   - If $Cov(X_1, X_2) = 0 \rightarrow$ unbiased coefficient, biased std. errors.

Misspecification: types of misspecifications

1. Omitted variables
   - True model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \varepsilon_i$
   - Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
   - If $Cov(X_1, X_2) \neq 0 \rightarrow$ biased coefficients and std. errors.
   - If $Cov(X_1, X_2) = 0 \rightarrow$ unbiased coefficient, biased std. errors.
2. Inclusion of irrelevant variables
   - True model: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
   - Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \varepsilon_i$
   - if $Cov(X_1, X_2) \neq 0 \rightarrow$ larger variances/inefficient.

Misspecification: types of misspecifications

1. Omitted variables
   - True model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \varepsilon_i$
   - Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
   - If $Cov(X_1, X_2) \neq 0 \rightarrow$ biased coefficients and std. errors.
   - If $Cov(X_1, X_2) = 0 \rightarrow$ unbiased coefficient, biased std. errors.
2. Inclusion of irrelevant variables
   - True model: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
   - Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_1 X_{2i} + \varepsilon_i$
   - if $Cov(X_1, X_2) \neq 0 \rightarrow$ larger variances/inefficient.
3. Incorrect functional form (non-linearity)
   - True model: $Y_i = \beta_0 + \beta_1 \log X_{1i} + \varepsilon_i$ or
     $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i$
   - Estimated model: $Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$
   - biased coefficients and std. errors.

Misspecification: dealing with omitted variable bias

Ex-post (if we cannot determine how the data is gathered):

- Use theory.
- Test for misspecification (ex. F-test).
- Use causal inference techniques:
  - Regression discontinuity.
  - Difference-in-differences.
  - Instrumental variables.

Ex-ante (if we can determine how the data is gathered).

- Randomization.

Misspecification: dealing with wrong functional form

Polynomial transformation:

- Polynomial relationship between $\hat{Y}_i$ and X_i.
- Think U-shaped or S-shaped relationships.
- Unlike logs, can reverse the direction of the relationship.
- See stata examples.
- Caution: interpretation of coefficients changes.

Types:

- **quadratic:** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{1i}^2$
- **cubic:** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{1i}^2 + \hat{\beta}_3 X_{1i}^3$

Misspecification: dealing with wrong functional form

Logarithmic transformation:

- Non-linear relationship between Y_i and X_i.
- Highly skewed variables.
- See stata examples.
- Caution: interpretation of coefficients changes.

Types:

- **linear-log:** $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \log X_i$
- **log-linear:** $\log \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- **log-log:** $\log \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \log X_i$

Misspecification: test

- Look for non-randomness in residual plots.
- RESET test.

Misspecification: RESET test

- Ramsey (1969)
- Purpose: test for non-linearities
- Pros: good test of misspecification.
- Cons: not a guide us as to alternative specifications.

1. Estimate your proposed: Ex. $Y = \beta_0 + \beta_1 X + u$
2. Compute fitted values $\hat{Y}$
3. Estimate: $Y = \beta_0 + \beta_1 X + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + v$
4. Test $H_0 : \delta_1 = \delta_2 = 0$
5. Perform a joint significance test on δ_1 and δ_2 ($F_{2,n-k-3}$).
6. Rejection of the null suggests existence of non-linearities.

Misspecification: F-test

- We might want to test a particular subset q of the coefficients.
- $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \cdots \beta_p = 0$
- H_1 : at least one $(\beta_1, ..., \beta_q) \neq 0$

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}$$

$RRS_0 =$ RRS for the model excluding the q parameters.

$RRS =$ RRS for the full (unrestricted) model.

Heteroskedasticity: consequences

- $Var(\varepsilon_i|X_i) \neq \sigma^2$ for some X_i
- Oftentimes $Var(\varepsilon_i)$ is a function of some X_i.
- $\rightarrow$ estimator of coefficient standard errors is biased.
- Problem for inference.
- Note: coefficient estimators remain unbiased but not efficient.
- One solution: use heteroskdasticity robust standard errors.

Heteroskedasticity: tests

1. Visual inspection:
   - Plot residuals as a function of the independent variables.
   - Plot squared residuals as a function of independent variables.
2. Goldfeld-Quandt Test.
3. White Test.

Heteroskedasticity: Goldfeld-Quandt test

1. Order obs according to X_i thought to be related to $Var(\varepsilon_i)$.
2. Take equally sized subsets of observations from both extremes.
3. Estimate model for each subset (ignoring the middle).
4. Perform F-test on the ratio of the residual sum of squares.
   - H_0: errors are homoskedastic.

Heteroskedasticity: White test

1. Estimate your proposed model: Ex. $Y = \beta_0 + \beta_1 X + u$
2. Compute residuals $\hat{e}$.
3. Regress $\hat{e}^2$ on the regressors, their squares and interactions.
4. Compute the chi-squared statistic $= n * R^2$
   - H_0: errors are homoskedastic.

Multicollinearity: consequences

- $Cov(X_i, X_j) \approx 1$ for some $X_i \neq X_j$
- Problem: large variances of the OLS parameter estimates.
- Note: OLS estimator still unbiased (indeed BLUE).
- Tradeoff: large parameter variances vs. omitted variable bias.

Multicollinearity: test

- Compute the variance inflation factor for each predictor.
- $VIF_j = \frac{1}{1-R_j^2}$
- $R_j^2 = R^2$ of regressing X_j on remaining predictors.
- Values > 10 suggest multicollinearity.

Multicollinearity: solutions

- Obtain more data.
  - Larger samples help reduce variance.
- Formalize relationship among regressors.
  - Simultaneous equation model.
- Drop a variable.
  - Tradeoff: large parameter variances vs. omitted variable bias.

Measurement error

- Measurement error on dep. variable: no problem.
  - Errors just become part of the disturbance term.
  - Bigger standard errors.
- Measurement error on indep. variable: problem if not random.
  - $Cov(\varepsilon_i, X_i) \neq 0$ for some X_i
  - Result: biased OLS estimator.
- Solutions:
  - Get better data.
  - Weighted regressions.
  - Instrumental variables.

Moderators and Mediators

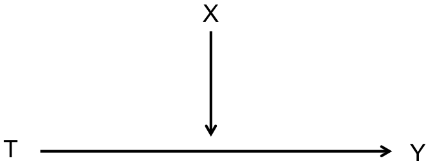

Figure 1: X as a moderator

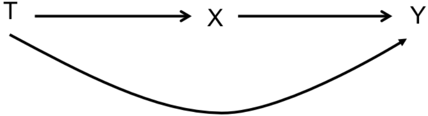

Figure 2: X as a mediator

Interaction Terms

Used when we believe the effect of T on Y is a function of X.

$\rightarrow$ X as a **moderator**.

We write:

$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 T * X + \varepsilon$$

For ease of interpretation rewrite as:

$$Y = \beta_0 + (\beta_1 + \beta_3 X) T + \beta_2 X + \varepsilon$$

$\rightarrow$ the effect of a unit change of T on $Y = (\beta_1 + \beta_3 X)$.

Marginal Effects

$$Y = \beta_0 + \beta_1 T + \beta_2 X + \beta_3 T * X + \varepsilon$$

Table 1: What does each of the coefficients represent?

	X_0 (X = 0)	X_1 (X = 1)	Difference (X_1 - X_0)
T_0 (T = 0)	β_0	$\beta_0 + \beta_2$	β_2
T_1 (T = 1)	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Difference (T_1 - T_0)	β_1	$\beta_1 + \beta_3$	β_3

Limited Dependent Variables

- Limited dependent variables (LDVs): outcome variables with finite, truncated, or discrete support.
- Example: $Y_i \in \{0, 1\}$. Here, the estimated effect β using OLS (i.e., a linear probability model) may be difficult to work with (functional form issues, inaccurate predictions).
- We need a model for $Pr[Y_i = 1|X_i]$, where X_i may be continuous or binary.

Limited Dependent Variables

- Begin by rescaling $E[Y_i|X_i] = Pr[Y_i = 1|X_i]$ as a linear function: $g(E[Y_i|X_i]) = X_i'\beta$.
- This gives us a *generalized linear model* (GLM), where $g(\cdot)$ is called the *link function* and $X_i'\beta$ is the linear predictor.
- The logistic transformation is one particular kind of link function which models *log-odds* (rather than probabilities).

Logistic Regression

log-odds scale

$$log\left(\frac{Pr[Y_i = 1|X_i]}{1 - Pr[Y_i = 1|X_i]}\right) = X_i\beta$$

- β's are only interpretable as signs $(+/-)$ and significance $(***)$.
- Must convert back to probability scale to get a substantive meaning of the coefficient. (Stata: - *margins* - command).

Instrumental Variables - Two Stage Least Squares (2SLS)

OLS

$$Y_i = \alpha + \beta D_i + \eta_i$$

IV: Second stage

$$Y_i = \alpha + \delta \widehat{D}_i + \varepsilon_i \tag{1}$$

IV: First stage

$$D_i = \alpha + \pi Z_i + \epsilon_i \tag{2}$$

Here, the instrument Z_i, for the treatment D_i, estimates the localized effect of the treatment on the outcome of interest Y_i.

Instrumental Variables - 2SLS

$$Cov[Y_i, Z_i] = Cov[\alpha + \beta D_i + \eta_i, Z_i] = \beta Cov[D_i, Z_i]$$

$$\Rightarrow \beta = \frac{Cov[Y_i, Z_i]}{Cov[D_i, Z_i]} = \frac{\frac{Cov[Y_i, Z_i]}{Var[Z_i]}}{\frac{Cov[D_i, Z_i}{Var[Z_i]}}$$

$$\Rightarrow \frac{\text{Reduced Form}}{\text{First stage}}$$

Where the 'Reduced Form' effect can be estimated using a regression: $Y_i = \alpha + \pi Z_i + \epsilon_{it}$

IV 2SLS Example: Haber et al., 2011 (APSR)

Second stage

$$Democracy_{it} = \alpha + \beta(\widehat{Oil}_{it}) + \delta_t + \gamma_i + \mathrm{X}'_{it}\lambda + \varepsilon_{it} \tag{3}$$

First stage

$$Oil_{it} = \alpha + \pi NaturalDisaster_{it} + \delta_t + \gamma_i + \mathrm{X}'_{it}\lambda + \epsilon_{it} \tag{4}$$

Strategy

Isolate source of *exogenous* variation (unforeseen natural disaster shocks) in oil producing countries to measure the causal effect of resource revenues on political institutions and human rights.

Instrumental Variables - Identification Assumptions

- The instrument is exogenous (i.e., it is as good as randomly assigned)[b].

- The instrument has some effect on regressor of interest (i.e., $E[D_{1i} - D_{0i}] \neq 0$). We can estimate the validity of the first-stage by examining the F-statistic ($F\text{-}stat > 10$ is good!).

- The "exclusion restriction" holds (i.e., the instrument has no effect on the outcomes of interest except through its effect on the regressor of interest D_i).

- The effects of the instrument on the regressor of interest are monotonic (i.e., $D_{1i} - D_{0i} \geq 0 \ \forall i$).

[b] More specifically, the instrument is orthogonal to the potential outcomes.