Regression analysis with two variables

Basic concepts (CH2)

Y: the expected consumption expenditures of a random household we want to determine
X_i: level of disposable income of that household
$E(Y|X_i) = \beta_1 + \beta_2 X_i$ **population regression function** (PRF) (linear both in parameters and variables), only on average correct ⇒ deviations presented as stochastic error term
$\mu_i = Y_i - E(Y|X_i)$
⇔ $Y_i = E(Y|X_i) + \mu_i = \beta_1 + \beta_2 X_i + \mu_i$
Error term contains all variables that affect Y but that are not included in the model
$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_i$ **sample regression function** (SRF)
If no data for an entire population, but only for a sample randomly taken, then there are estimators (parameters) needed to make an approximation, based on an estimator (method)
Stochastic: changing over repeated sampling
Deterministic: constant over repeated sampling

Estimating the sample regression function (CH3)

Ordinary least squares method (OLS) to avoid +/- errors cancelling out and $|\mu_i|$ less interesting

Numerical properties of the OLS estimator

1. Sample regression line passes through the sample means of Y and X
2. Mean $\widehat{Y}_i$ = mean Y_i
3. Average $\overline{\widehat{\mu}_i}$ = 0
4. $\widehat{\mu}_i$ not correlated with X_i
5. $\widehat{\mu}_i$ not correlated with Y_i

Gauss-Markov assumptions

Additional assumptions ⇒ classical linear regression model (CLRM), tied to PRF not SRF
1. Linearity in the parameters
2.
    a. X-values fixed over repeated sampling: **fixed regressor model**
    b. X-values changing over repeated sampling: **stochastic regressor model**
3. No **systematically** affection by variables/factors excluded from the model
    a. $E(\mu_i)$=0 (deterministic X_i)

b. $E(\mu_i|X_i)=0$ (stochastic X_i)

⇒ X_i and μ_i not correlated

4. Variance of μ_i constant (homoscedasticity)

   $var(\mu_i|X_i) = \sigma^2$ *violation:* $var(\mu_i|X_i) = \sigma_i^2$ (heteroscedasticity)
5. No correlation in error terms ⇒ no systematic pattern in error terms

   $cov(\mu_i,\mu_j|X_i,X_j) = 0$ for $i \neq j$ *violation:* autocorrelation
6. #observations > #parameters to be estimated
7. Variation in X-values
8. No perfect multicollinearity

Precision of the OLS estimator

~variability of $\widehat{\beta_1}$ and $\widehat{\beta_2}$ over repeated sampling ⇒ $se(\widehat{\beta_1})$ and $se(\widehat{\beta_2})$

standard error:

Interpretation: standard deviation of the sampling distribution of this estimator

Estimation: $se(\widehat{\beta}) \rightarrow \widehat{se}(\widehat{\beta})$ formulas on sheet

$\text{var}(\hat{\beta}_2) = \frac{1}{\sum x_i^2}\sigma^2$ $\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2$ calculate the true variances, but μ_i ⇒ σ not known

$\hat{\sigma} = \sqrt{\frac{\sum \hat{\mu}_i^2}{n-2}}$ estimator for σ² and σ, **unbiased:** $E(\widehat{\sigma}^2) = \sigma^2$, n-2 degrees of freedom

σ = standard error of the regression

The bigger the variance of the residuals, the lower the precision

The bigger the variance of the explanatory variable or the bigger the samplesize, the higher the precision

Statistical properties of the OLS estimator

All GM assumptions satisfied ⇒ OLS = best linear unbiased estimator (BLUE)

Linear: estimator linear function of stochastic Y

Unbiased: expected value $E(\widehat{\beta}) = \beta$ the true population

Efficient: smallest variance

Coefficient of determination (R²)

Measures how well the estimated regression line fits the sample data

Calculates the proportion of the variance of Y_i explained by the variance of X_i

The normality assumption (CH4)

Extra assumptions needed because OLS never gives β_1 and β_2 ⇒ hypothesis testing
Probability distribution of the parameters needed
X_i deterministic ⇒ $\widehat{\beta}_2$ weighted average of μ_i with weight k_i fixed over repeated sampling

$\widehat{\beta}_2 = \beta_2 + \sum k_i \mu_i$

Classical linear regression model assumes μ_i normally distributed (homoscedasticity)
CLRM ⇒ CNLRM
$Y_i = \beta_1 + \beta_2 X_i + \mu_i$ where $\mu_i \sim NID(0, \sigma^2)$ normally independently distributed (no autocorrelation)

Properties of OLS estimator

1. $\widehat{\beta}_1$, $\widehat{\beta}_2$ unbiased, efficient, normally distributed
2. $\widehat{\sigma}^2$ unbiased and χ^2 distributed
3. $\widehat{\beta}_1$ and $\widehat{\beta}_2$ distributed independently of $\widehat{\sigma}^2$

⇒ $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are BUE: efficient for entire set of unbiased estimator (linear and non-linear)
⇒ $Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$ because linear function of deterministic X_i

Interval estimation and hypothesis testing (CH5)

Interval estimation: adding a margin to estimator such that it contains the true population with a certain probability
$Pr(\widehat{\beta} - \delta \le \beta \le \widehat{\beta} + \delta) = 1 - \alpha$ with $0 < \alpha < 1$
α significance level, $1 - \alpha$ confidence coefficient (stochastic)
Interpretation: when constructing confidence intervals with a confidence coefficient $1 - \alpha$, over repeated sampling these intervals will contain the true population parameter β in ($1 - \alpha$)% of the cases
Population parameter σ^2 is unknown ⇒ use $\widehat{\sigma}^2$ as an unbiased estimator
Which leads to confidence intervals
$Pr(\widehat{\beta}_1 - t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\beta}_1} \le \beta_1 \le \widehat{\beta}_1 + t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\beta}_1}) = 1 - \alpha$
$Pr(\widehat{\beta}_2 - t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\beta}_2} \le \beta_2 \le \widehat{\beta}_2 + t_{n-2,\alpha/2}\widehat{\sigma}_{\widehat{\beta}_2}) = 1 - \alpha$
Hypothesis testing: formulate a hypothesis $\beta_2 = \beta_2^*$, check whether is is possible by checking whether $\widehat{\beta}_2$ is sufficiently close to β_2^* using the statistical properties of the OLS estimator

Two-sided hypothesis

$H_0 : \beta_2 = \beta_2^*$ proposed hypothesis = null hypothesis
$H_1 : \beta_2 \neq \beta_2^*$ alternative hypothesis

One-sided hypothesis

$H_0 : \beta_2 \leq \beta_2^*$ or $H_0 : \beta_2 \geq \beta_2^*$
$H_1 : \beta_2 > \beta_2^*$ $H_1 : \beta_2 < \beta_2^*$

Via confidence interval: reject H_0 if β_2^* does not lie within the confidence interval
Via significance test: compute test statistic under H_0: $\beta_2 = \beta_2^*$ ⇒ $t = \frac{(\hat{\beta}_2 - \beta_2^*)}{\hat{\sigma}}$
and reject H_0 if $|t| > t_{n-2,\alpha/2}$

Terminology

Statistically significant: if H_0 can be rejected
If t-test not significant, H_0 cannot be rejected
Never accept, only reject or don't reject
type-I error: rejecting H_0 while correct, probability upper limit α **'size'** as low as possible
type-II error: not rejecting H_0 while incorrect, probability β ⇒ $1 - \beta$ **'power'** as low as possible
⇒ trade-off

One-sided hypothesis test

If strong a priori indications
H_0 gets benefit of the doubt ⇒ theoretical proposition under the alternative hypothesis

The '2-t' rule of thumb

If test statistic > 2 reject H_0 because $t = \hat{\beta}_2 / \hat{\sigma}_{\hat{\beta}_2} = 1,96 \approx 2$

Exact significance level (p-value)

Lowest point at which H_0 can be rejected
Exact probability of making a type-I error
No information about the power

Analysis of variance (ANOVA)

H_0: $\beta_2 = 0$ all of the variance in Y results from variance in μ (ESS=0)
H_1: $\beta_2 \neq 0$ a part of the variance in Y results from variance in X (ESS>0)

Relation between t and F-test for k_1 = 1

$\sqrt{F} \approx t$

Some extensions (CH6)

Interpreting regression results

$Y_i = \beta_1 + \beta_2 X_i + \mu_i$

β_1 intercept: expected level of Y_i when X_i=0

β_2 slope: expected change in Y_i when X_i increases by 1

e_Y^X elasticity: an increase in X_i by 1 percent induces an expected change in Y_i by $\beta_2 X_i / Y_i$ percent

Linear in logs model

Consider an **exponential model**

$$Y_i = \beta_1 X_i^{\beta_2} e^{\mu_i}$$

Using a **logarithmic transformation**

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + \mu_i$$

and setting $\alpha = \ln \beta_1$, $Y_i^* = \ln Y_i$ and $X_i^* = \ln X_i$

$$Y_i^* = \alpha + \beta_2 X_i^* + \mu_i$$

This model can be estimated using OLS, because it is linear

- in the parameters (α, β_2)
- in de transformed variabels (Y_i^*, X_i^*)

Properties: under CNLRM ⇒ X_i deterministic, $\mu_i \sim NID(0, \sigma^2)$, OLS estimators $\widehat{\alpha}$ and $\widehat{\beta}_2$ BUE

Interpretation: β_2 measures elasticity of Y_i to changes in X_i

Semi-log model

Consider **log-lin model**

$$\ln Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

Setting $Y_i^* = \ln Y_i$

$$Y_i^* = \beta_1 + \beta_2 X_i + \mu_i$$

Interpretation: for a one unit absolute change in X_i

- β_2 measures the relative change in Y_i

$$\beta_2 = \frac{\Delta Y_i^*}{\Delta X_i} = \frac{\Delta \ln Y_i}{\Delta X_i} \approx \frac{\Delta Y_i / Y_i}{\Delta X_i}$$

- $100\beta_2$ measures the percentage change in Y_i

$$100\beta_2 = \frac{100\Delta Y_i / Y_i}{\Delta X_i}$$

Consider **lin-log model**

$$Y_i = \beta_1 + \beta_2 \ln X_i + \mu_i$$

Setting $X_i^* = \ln X_i$

$$Y_i = \beta_1 + \beta_2 X_i^* + \mu_i$$

Interpretation: the absolute change in Y_i is measured by

- β_2 for a relative change in X_i

$$\beta_2 = \frac{\Delta Y_i}{\Delta X_i^*} = \frac{\Delta Y_i}{\Delta \ln X_i} \approx \frac{\Delta Y_i}{\Delta X_i / X_i}$$

- $\beta_2 / 100$ for a percentage change in X_i

$$\beta_2 / 100 = \frac{\Delta Y_i}{100\Delta X_i / X_i}$$

Reciprocal model

Consider the following model

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + \mu_i$$

Interpretation:

- β_1 is is the asymptotic value for Y_i when $X_i \to \infty$
- β_2 measures

$$\beta_2 = \frac{\Delta Y_i}{\Delta (1/X_i)} = \frac{\Delta Y_i / \Delta X_i}{\Delta (1/X_i)/\Delta X_i} \approx -\frac{\Delta Y_i / \Delta X_i}{(1/X_i)^2}$$

$$\to \Delta Y_i = -\beta_2 \frac{\Delta X_i}{X_i^2}$$

Choice of functional form

Ideally theory based, but alignment with data mandatory ⇒ check R^2 but only if same dependent variable (so no Y_i with $\ln(Y_i)$)

Multivariate regression analysis

Estimating the sample regression function (CH7)

For OLS to be unbiased, $E(\mu_i) = 0$, all relevant variables have to be included in the model
If one variable (A) is strongly correlated with a variable (B) and another one (C), then the OLS parameter for C may include impact of A on B
⇒ A is a **confounding variable**, which needs to be controlled when estimating the impact of C on B

Notation and interpretation

The population regression curve is the **locus of the conditional expectations** of Y_i for fixed values of X_{2i} & X_{3i}:

$$E(Y_i | X_{2i}, X_{3i}) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i}$$

β_2 and β_3 are **partial regression coefficients** (ceteris paribus)

$$\beta_2 = \frac{\partial E(Y_i | X_{2i}, X_{3i})}{\partial X_{2i}}, \qquad \beta_3 = \frac{\partial E(Y_i | X_{2i}, X_{3i})}{\partial X_{3i}}$$

- β_2 indicates the change in $E(Y_i | X_{2i}, X_{3i})$ for $\Delta X_{2i} = 1$ and $\Delta X_{3i} = 0$, i.e. the <u>direct or net impact of X_{2i} on Y_i</u>
- β_3 indicates the change in $E(Y_i | X_{2i}, X_{3i})$ for $\Delta X_{3i} = 1$ and $\Delta X_{2i} = 0$, i.e. the <u>direct or net impact of X_{3i} on Y_i</u>

Even when we are only interested in the direct impact β_2 of X_{2i} we need to include X_{3i} as a **control variable**

When estimating the following model:

$$Y_i = \alpha_1 + \alpha_2 X_{2i} + \varepsilon_i$$

the OLS estimator $\widehat{\alpha}_2$ is (in general) a biased and inconsistent estimator of β_2 in

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i$$

- β_2 measures **direct impact** of X_{2i} on Y_i (i.e. for X_{3i} fixed)
- α_2 also captures **part of the impact of** X_{3i} (i.e. X_{3i} is allowed to change and may be correlated with X_{2i})

Least squares estimation

Orthogonal projection (two-step approach):

Step 1

Impact of FLR can be eliminated from CM by regressing CM on FLR using OLS

$$CM_i = b_1 + b_{1,3} FLR_i + \mu_{1i}$$

and save the estimated residuals $\widehat{\mu}_{1i}$, with $\text{cov}\left(FLR_i, \widehat{\mu}_{1i}\right) = 0$ (numerical property OLS!)

Impact of FLR can be eliminated from $PGNP$ by regressing $PGNP$ on FLR using OLS

$$PGNP_i = b_2 + b_{2,3} FLR_i + \mu_{2i}$$

and save the estimated residuals $\widehat{\mu}_{2i}$, with $\text{cov}\left(FLR_i, \widehat{\mu}_{2i}\right) = 0$ (numerical property OLS!)

Step 2

Regressing $\widehat{\mu}_{1i}$ on $\widehat{\mu}_{2i}$

$$\widehat{\mu}_{1i} = a_1 \widehat{\mu}_{2i} + \mu_{3i}$$

yields a_1 as an estimator for the partial regression coefficient β_2 in the original multivariate model

$$CM_i = \beta_1 + \beta_2 PGNP_i + \beta_3 FLR_i + \mu_i$$

$a_1 = \widehat{\beta}_2 = -0.0056$ in child mortality example

Multivariate ordinary least squares (one-step approach):

Parameters obtaining using LS criterion

$$\min_{\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3} \sum \widehat{\mu}_i^2 = \min_{\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3} \sum \left(Y_i - \widehat{\beta}_1 - \widehat{\beta}_2 X_{2i} - \widehat{\beta}_3 X_{3i} \right)^2$$

from which three **first order conditions** can be derived

$$1. -2 \sum \left(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \right) = 0 \quad \rightarrow \sum \hat{\mu}_i = 0$$

$$2. -2 \sum X_{2i} \left(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \right) = 0 \quad \rightarrow \sum X_{2i} \hat{\mu}_i = 0$$

$$3. -2 \sum X_{3i} \left(Y_i - \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} \right) = 0 \quad \rightarrow \sum X_{3i} \hat{\mu}_i = 0$$

Which can be uses to calculate the 3 parameters (see formula sheet)

Numerical properties of the OLS estimator

$$\overline{Y} = \widehat{\beta}_1 + \widehat{\beta}_2\overline{X}_2 + \widehat{\beta}_3\overline{X}_3$$

$$\overline{\widehat{Y}} = \overline{Y}$$

$$\frac{1}{n}\sum \widehat{\mu}_i = \overline{\widehat{\mu}} = 0$$

$$\frac{1}{n}\sum \widehat{\mu}_i X_{2i} = \frac{1}{n}\sum \widehat{\mu}_i X_{3i} = 0$$

$$\frac{1}{n}\sum \widehat{\mu}_i \widehat{Y}_i = 0$$

Statistical properties of the OLS estimator

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i$ where $\mu_i \sim NID(0, \sigma^2)$

Regularity conditions:

- X_{2i} and X_{3i} deterministic
- #observations n > #parameters to be estimated k
- Positive variance in X_{2i} and X_{3i}
- No perfect multicollinearity: $|cor(X_{2i}, X_{3i})| = |r_{23}| \neq 1$

⇒ $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are BUE and normally distributed

Polynomial regression models

Quadratic specification (second order polynomial)

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2$$

Stochastic specification

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \mu_i$$

More general: a k-th order polynomial is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \ldots + \beta_k X_i^k + \mu_i$$

Interpretation

$$\frac{\partial Y_i}{\partial X_i} = \beta_1 + 2\beta_2 X_i + 3\beta_2 X_i^2 + \ldots + k\beta_k X_i^{k-1}$$

Multivariate determination coefficient R²

Indicates how well the estimated regression line fits the data by calculating the part of the variance in Y_i that is explained by the variance in X_{2i} and X_{3i}

Adjusted determination coefficient R²

R² always increases when explanatory variables are added ⇒ not useful when comparing 2 models with different number of explanatory variables
ModifiedR² = (1-k/n)R²

Relation between R^2 and $\overline{R}^2$

$$\overline{R}^2 = 1 - \left(1 - R^2\right) \frac{n-1}{n-k}$$

- $\overline{R}^2 < R^2$ for $k > 1$
- $\overline{R}^2$ can be negative, i.e. for $R^2 < (k-1)/(n-1)$

Normality assumption and hypothesis testing (CH8)

$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \mu_i$ where $\mu_i \sim NID(0, \sigma^2)$
$\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ are BUE and normally distributed

Statistical properties of OLS estimator

$\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ normally distributed
$(n-3)\widehat{\sigma}^2/\sigma^2 \sim \chi^2_k$ where k = df = n - 3
$\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$ independently distributed from $\widehat{\sigma}^2$
$t = (\widehat{\beta}_i - \beta_i)/\widehat{\sigma}_{\widehat{\beta}_i} \sim t_{n-3}$

Joint significance of all coefficients

ANOVA: H_0: $\beta_2 = \beta_3 = 0$
F-test: reject when > critical value
Relation with R²: R²=0 ⇒ F=0, R²=1 ⇒ F=∞

Significance marginal contribution

Sequential regression: ANOVA check whether adding variables change the explanatory power (ESS) significantly
Remark: changing the order in which variables are added has an impact on the results, because adding A highly collinear with B that is already in the model will not impact the ESS a lot (thus concluding B only significant variable), switching the variables ⇒ A only significant

General procedure F-test

Test linear restrictions H_0
Estimate 'general', unrestricted model, compute RSS_{UR}
Derive the constrained model by imposing H_0
Estimate constrained model, calculate RSS_R
Test the hypothesis using F-test

$$F = \frac{(\text{RSS}_\text{R} - \text{RSS}_\text{UR})/m}{\text{RSS}_\text{UR}/(n-k)} \sim F_{m,(n-k)}$$

Chow test: coefficient stability

Problem: time series ⇒ not stable over time, cross-sectional ⇒ not stable over groups/units

Assumptions:

- $\mu_{1,t} \sim N(0, \sigma^2)$ and $\mu_{2,t} \sim N(0, \sigma^2)$
- $\mu_{1,t}$ en $\mu_{2,t}$ are independently distributed

Procedure:

- Estimate the different models and calculate RSS_1, RSS_2, RSS_3
- Calculate $RSS_{UR} = RSS_1 + RSS_2$ en $RSS_R = RSS_3$
- Calculate F-statistic

$$F = \frac{(RSS_R - RSS_{UR})/k}{RSS_{UR}/(n_1 + n_2 - 2k)} \sim F_{k,(n_1+n_2-2k)}$$

Limitations:

Variance of the error terms has to be constant over sub-periods (homoskedasticity)

Test does not tell us whether rejection of H_0 is due to instability in the intercept or in the slope (see chapter 9)

Break point has to be known

Regression with dummy variables (CH9)

Necessary when variables are qualitative or categorical

Consequences for OLS

If qualitative as explanatory ⇒ model is linear ⇒ OLS appropriate
If qualitative as dependent ⇒ typically non-linear model ⇒ OLS inappropriate, estimation using maximum likelihood (ML)
Reference category: for which no dummy is included, choice does not influence results
Example: 0 = black, 1 = white ⇒ black is reference category

Dummy variable trap

When m explanatory qualitative variables, only m-1 dummies can be included in a model with a constant
Otherwise perfect multicollinearity

Relaxing the assumptions of CLNRM

Multicollinearity (CH10)

(Perfect) multicollinearity: (perfect) linear relation between some or all explanatory variables
Causes: dummy variables, model specification (X_i and X_i^2), large number of explanatory variables, lack of data…
Consequences: parameters can't be estimated (perfect mc), estimators have larger (co)variance, wider confidence intervals and lower t-stats
It's a sample problem: highly correlated variables, too much variance filtered out in multivariate regression
Though OLS still unbiased and efficient (both only over repeated sampling), coefficients can't be estimated precisely

Detection

= measuring the degree of multicollinearity using rules-of-thumb
Compare R^2 with t-values (high with low)
Calculate pairwise correlation (high)
Estimate auxiliary regressions
Compute variance inflation factor (VIF) for each variable (high compared to samplesize)

Remedial measures

Cannot be solved by changing estimation method
Richer dataset
Adjust specification

Heteroskedasticity (CH11)

Variance of μ_i is not constant
Causes: population (error learning), data collection (outliers), specification errors (dropping relevant variables, wrong functional form)
Consequences: OLS no longer efficient ⇒ GLS lower variance

Generalized least squares (GLS)

Assume

$$Y_i = \beta_1 X_{0i} + \beta_2 X_i + \mu_i$$

where $X_{0i} = 1 \ \forall i$ and $E\left(\mu_i^2\right) = \sigma_i^2$ are known

- Transform the model by dividing by σ_i

$$\frac{Y_i}{\sigma_i} = \beta_1 \frac{X_{0i}}{\sigma_i} + \beta_2 \frac{X_i}{\sigma_i} + \frac{\mu_i}{\sigma_i}$$
$$Y_i^* = \beta_1 X_{0i}^* + \beta_2 X_i^* + \mu_i^*$$

- $\operatorname{var}\left(\mu_i^*\right) = E\left(\mu_i^{*2}\right) = E\left(\mu_i^2 / \sigma_i^2\right) = E\left(\mu_i^2\right) / \sigma_i^2 = 1$
- μ_i^* is homoskedastic
- Assumptions CNLRM are fulfilled for the transformed model
- OLS on the transformed model (=GLS) is **BUE and normally distributed**
- OLS on original model is not efficient: $\operatorname{var}\left(\widehat{\beta}_2^{GLS}\right) \leq \operatorname{var}\left(\widehat{\beta}_2^{OLS}\right)$

Intuition for the efficiency of GLS

- OLS minimizes unweighted sum: $\sum \widehat{\mu}_i^2$
- GLS minimizes weighted sum: $\sum w_i \widehat{\mu}_i^2$ with $w_i = 1 / \sigma_i^2$
  - more weight is given to observations for which we expect that they will be closer to the population regression curve, i.e. for which the variance in the error terms is smaller [Fig. 11.7]
  - alternative name: **weighted least squares** (WLS)

Consequences for testing based on OLS

Heteroscedasticity acknowledged: valid inference, but wider confidence interval, lower significance
Heteroskedasticity ignored: invalid inference, estimator biased and inconsistent

Detection

Calculate σ_i^2 for entire population, check whether constant
In practise: use estimator $\widehat{\mu}_i^2$ for σ_i^2, check whether constant
Informal: intuitive(in cross-sections, heteroskedasticity rule rather than exception), graphical
Formal: Goldfeld-Quandt test (non parametric), White's general heteroscedasticity test (parametric)

Goldfeld-Quandt test

Assumption: σ_i^2 positively related to one of the explanatory variables
Test procedure: order observations based on X_i, delete c (=⅙ of pop.) obs in the middle apply OLS to the 2 groups
Under $\mu_i \sim N$ and under $H_0 : \sigma_1^2 = \sigma_2^2$ the following holds:

$$\lambda = \frac{\widehat{\sigma}_2^2 / \sigma_2^2}{\widehat{\sigma}_1^2 / \sigma_1^2} = \frac{\widehat{\sigma}_2^2}{\widehat{\sigma}_1^2} \sim F_{(n-c)/2-k,(n-c)/2-k}$$

or

$$\lambda = \frac{\frac{RSS_2}{(n-c)/2-k}}{\frac{RSS_1}{(n-c)/2-k}} = \frac{RSS_2}{RSS_1} \sim F_{(n-c)/2-k,(n-c)/2-k}$$

Reject H_0 of homoskedasticity if λ is larger than critical value

White’s general heteroscedasticity test

Test procedure:

1. Estimate the model and calculate $\widehat{\mu}_i$
2. Estimate the following auxiliary regression

$$\widehat{\mu}_i^2 = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{2i}^2 + \alpha_5 X_{3i}^2 + \alpha_6 X_{2i} X_{3i} + \nu_i$$

3. H_0: μ_i is homoskedastic $\rightarrow \alpha_2 = \alpha_3 = \ldots = \alpha_6 = 0$
4. White: under H_0 and as $n \rightarrow \infty$

$$nR^2 \sim \chi^2_{df}$$

   where df is equal to the number of explanatory variables (excluding the constant) in the auxiliary regression

Homoscedasticity: $R^2=0$

Remarks

- In principle, H_0 can be tested using standard F-test, but exact small sample distribution is unknown
- Large drop in degrees of freedom in regressions including many explanatory variables
  - Test can be applied without cross products
- A significant test statistic may also be due to specification error (e.g. model not linear in the variables)

Remedial measures

Check slides

Autocorrelation (CH12)

There is a systematic pattern in the error terms, positive or negative
Mostly relevant for time series and panel data
Causes: inertia, transformation of the data, specification errors, non-stationarity
Consequences: assumption ⇒ μ_t follows an AR(1) process = autoregressive of the first order, OLS no longer efficient ⇒ GLS

Properties AR(1) process

Backward iteration:

$$
\begin{aligned}
\mu_t &= \rho\mu_{t-1} + \varepsilon_t \\
&= \rho\left(\rho\mu_{t-2} + \varepsilon_{t-1}\right) + \varepsilon_t \\
&= \rho^2\mu_{t-2} + \varepsilon_t + \rho\varepsilon_{t-1} \\
&= \rho^2\left(\rho\mu_{t-3} + \varepsilon_{t-2}\right) + \varepsilon_t + \rho\varepsilon_{t-1} \\
&= \rho^3\mu_{t-3} + \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} \\
&= \ldots \\
&= \rho^t\mu_0 + \varepsilon_t + \rho\varepsilon_{t-1} + \rho^2\varepsilon_{t-2} + \ldots + \rho^{t-1}\varepsilon_1
\end{aligned}
$$

For $t \to \infty$ (since $|\rho| < 1$)

$$\mu_t = \sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}$$

Expected value

$$E\left(\mu_t\right) = E\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}\right) = \sum_{i=0}^{\infty} \rho^i E\left(\varepsilon_{t-i}\right) = 0$$

Variance

$$
\begin{aligned}
\mathrm{Var}\left(\mu_t\right) = E\left(\mu_t^2\right) &= E\left(\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}\right)^2\right) \\
&= \sum_{i=0}^{\infty} \rho^{2i} E\left(\varepsilon_{t-i}^2\right) = \sigma^2 \sum_{i=0}^{\infty} \rho^{2i} = \frac{\sigma^2}{1-\rho^2}
\end{aligned}
$$

Covariance

$$
\begin{aligned}
\mathrm{Cov}\left(\mu_t \mu_{t-s}\right) &= E\left(\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-i}\right)\left(\sum_{i=0}^{\infty} \rho^i \varepsilon_{t-s-i}\right)\right) \\
&= \sum_{i=0}^{\infty} \rho^{s+2i} E\left(\varepsilon_{t-s-i}^2\right) = \sigma^2 \rho^s \sum_{i=0}^{\infty} \rho^{2i} \\
&= \rho^s \frac{\sigma^2}{1-\rho^2}
\end{aligned}
$$

Correlation

$$\mathrm{Cor}\left(\mu_t \mu_{t-s}\right) = \rho^s$$

Detection

Runs test (nonparametric): #runs R outside confidence interval
Durbin Watson d test: d close to 0 or 4

Assumptions: X_i deterministic, AR(1) pattern, μ_i normally distributed, no lagged dependent variables like Y_{t-1}

Breusch-Godfrey LM test: $\rho_i \neq \rho_j$ with i ≠ j

If specifications errors present, model becomes inconsistent ⇒ other tests

Dynamic models

Model sluggish reaction of Y_t to 'impulses'

- Autoregressive model: add lagged dependent variable Y_{t-1}

$$Y_t = \alpha + \rho Y_{t-1} + \beta_2 X_t + \varepsilon_t$$

- Distributed lag model: add lagged explanatory variable X_{t-1}

$$Y_t = \alpha + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

- Autoregressive Distributed Lag (ADL) model

$$Y_t = \alpha + \rho Y_{t-1} + \beta_2 X_t + \beta_3 X_{t-1} + \varepsilon_t$$

- Add deeper lags: ADL(1, 1) → ADL(p, q)

Possible approach: add lags until autocorrelation in ε_t is removed

Note the similarity and difference between (E)GLS and ADL

- Starting from assumption of AR(1) error terms (similar for higher order models)
- GLS transformation implies

$$Y_t - \rho Y_{t-1} = \beta_1 - \rho\beta_1 + \beta_2 X_t - \rho\beta_2 X_{t-1} + \mu_t - \rho\mu_{t-1}$$
$$Y_t = \alpha + \rho Y_{t-1} + \beta_2 X_t - \rho\beta_2 X_{t-1} + \varepsilon_t$$

- GLS imposes a non-linear restriction ($\beta_3 = -\rho\beta_2$) on the ADL(1, 1) model
- Standard t and F-tests not possible (but more complex alternatives available)
- Pragmatic approach: check autocorrelation in the error terms (as this is where and how a specification error should show up)

Specification errors (CH13)

Nature of problem: exclude relevant/ include irrelevant variable, wrong functional form, measurement errors

Consequences:

Excluding relevant variable

1. OLS is **biased and inconsistent** when $r_{23} = cor(X_{2i}, X_{3i}) \neq 0$

$$E(\widehat{\alpha}_2) = \beta_2 + b_{32}\beta_3$$

where $b_{32} = \sum x_{2i}x_{3i} / \sum x_{2i}^2$ (see App 13A.1)

2. Estimator variance error terms

$$E(\widehat{\sigma}_\nu^2) = \sigma_\nu^2 \neq \sigma_\mu^2$$

3. Variance OLS estimator

$$\mathrm{Var}(\widehat{\alpha}_2) = \frac{\sigma_\nu^2}{\sum x_{2i}^2} \neq \mathrm{Var}\left(\widehat{\beta}_2\right) = VIF \frac{\sigma_\mu^2}{\sum x_{2i}^2}$$

- $\mathrm{Var}(\widehat{\alpha}_2) < \mathrm{Var}\left(\widehat{\beta}_2\right)$ when $\sigma_\nu^2 / \sigma_\mu^2 < VIF$ $(r_{23}$ large, $\sigma_\nu^2 \approx \sigma_\mu^2)$
- $\mathrm{Var}(\widehat{\alpha}_2) > \mathrm{Var}\left(\widehat{\beta}_2\right)$ when $\sigma_\nu^2 / \sigma_\mu^2 > VIF$ $(r_{23}$ small, $\sigma_\nu^2 > \sigma_\mu^2)$

Impact on variance even when $r_{23} = 0$!!!

Including irrelevant variable ⇒ lose accuracy

1. OLS estimator is **unbiased and consistent** (see App 13A.2)

$$E(\widehat{\alpha}_2) = \beta_2$$
$$E(\widehat{\alpha}_3) = \beta_3 = 0$$

2. Estimator variance error terms

$$E(\widehat{\sigma}_\nu^2) = \sigma_\nu^2 = \sigma_\mu^2$$

3. Variance OLS estimator

$$\mathrm{Var}(\widehat{\alpha}_2) = VIF \frac{\sigma_\nu^2}{\sum x_{2i}^2} = VIF \frac{\sigma_\mu^2}{\sum x_{2i}^2} \geq \frac{\sigma_\mu^2}{\sum x_{2i}^2}$$

since $VIF \geq 1$

Wrong functional form

OLS estimator biased and inconsistent

Measurement errors in dependent variable

1. KK is **unbiased and consistent**

$$E\left(\widehat{\beta}_2^*\right) = \beta_2$$

2. Estimator variance error terms

$$E\left(\widehat{\sigma}_\nu^2\right) = \sigma_\nu^2 = \mathrm{Var}\left(\mu_i + e_{1i}\right) = \sigma_\mu^2 + \sigma_e^2$$

3. **Variance OLS estimator increases**

$$\mathrm{Var}\left(\widehat{\beta}_2^*\right) = \frac{\sigma_\mu^2 + \sigma_e^2}{\sum x_{2i}^2} > \mathrm{Var}\left(\widehat{\beta}_2\right) = \frac{\sigma_\mu^2}{\sum x_{2i}^2}$$

Measurement errors in explanatory variable

1. OLS is **biased and inconsistent** since

   $$E\left(X_i^* \nu_i\right) = E\left(\left(X_i + e_i\right)\left(\mu_i - \beta_2 e_i\right)\right) = -\beta_2 E\left(e_i^2\right) = -\beta_2 \sigma_e^2$$

   This is an **endogeneity** issue

   $$\rightarrow \mathrm{plim}\left(\widehat{\beta}_2\right) = \beta_2 \frac{1}{1 + \sigma_e^2 / \sigma_X^2} \qquad \text{(see App. 13A.3)}$$

   $$\mathrm{plim}\left(\widehat{\beta}_2\right) \leq \beta_2$$

2. Estimator variance error terms

$$E\left(\widehat{\sigma}_\nu^2\right) = \sigma_\nu^2 = \mathrm{Var}\left(\mu_i - \beta_2 e_i\right) = \sigma_\mu^2 + \beta_2^2 \sigma_e^2$$

3. **Variance OLS estimator increases**

$$\mathrm{Var}\left(\widehat{\beta}_2^*\right) = \frac{\sigma_\mu^2 + \beta_2^2 \sigma_e^2}{\sum x_{2i}^2} > \mathrm{Var}\left(\widehat{\beta}_2\right) = \frac{\sigma_\mu^2}{\sum x_{2i}^2}$$

Model selection criteria

Be consistent with the theory and data, encompassing, establish causality
Purist: specification based on theory, then check whether irrelevant variables are included, general-to-specific approach
⇒ often too strict
Data mining: only variables that are significant are tested and added, specific-to-general, to fit the data as well as possible
⇒ risk: significant correlations but not necessarily causal relations, real significance no longer nominal significance level
Solution: set part of sample aside, compute a next observation, compare to observation you set aside and check whether model fits the data

Detection

Ramsey's RESET test: add non-linear transformations of $\widehat{Y}_i$ to Y_i, use F-test, rejecting H_0 ⇒ specification error
Lagrange Multiplier (LM) test: add non-linear transformations of X_i to $\widehat{\mu}_i$, compute nR^2, reject H_0 if > critical value
Forecast χ^2 test: use one part to estimate, use the other part to test the out-of-sample performance

Endogeneity (CH18-20)

Reverse causality: simultaneous equation model with at least 2 endogenous variables that influence each other
Consequences: X_i stochastic and X_i and μ_i both dependent ⇒ OLS inconsistent

$$\text{plim} \sum k_i \mu_i = \frac{\text{plim} \frac{1}{n} \sum x_i \mu_i}{\text{plim} \frac{1}{n} \sum x_i^2} = \frac{E(\mu_i x_i)}{E(x_i^2)} \neq 0$$