Study Guide for Econometrics

Unit 1: Ordinary Least Squares Regression

OLS Model: \( y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + e_i \) (Wooldridge 2.1)

(Univariate: \( y_i = \beta_1 + \beta_2 x_i + e_i \))

Econometric model: \( \hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{2i} + \ldots + \hat{\beta}_k x_{ki} + \hat{\beta}_1 x_{ki} + \hat{e}_i \).

OLS Estimation (Wooldridge 2.2)

Objective: pick \( \hat{\beta}_{OLS} \) to minimize \( \sum_{i=1}^{N} \hat{e}_i^2 \).

OLS estimator: \( \hat{\beta}_2^{OLS} = \text{cov}(x_i, y_i) / \text{var}(x_i) \); \( \hat{\beta}_1^{OLS} = \bar{y} - \hat{\beta}_2^{OLS} \bar{x} \)

Expected value: if \( E[e_i | x_{2i}, \ldots, x_{ki}] = 0 \), then \( E[\hat{\beta}_{OLS}] = \beta \).

Estimated variance: \( \text{var}(\hat{\beta}_2^{OLS}) = \sum \hat{e}_i^2 / (N - K) \text{var}(x_i) \)

Standard errors: square root of estimated variance

If \( e_i \sim N(0, \sigma^2_e) \), then \( \hat{\beta}_{OLS} \) is normal. (Otherwise, approximately normal in large samples.)

Efficiency: If \( \text{var}(e_i) = \text{var}(e_j) = \sigma^2_e \) and \( \text{cov}(e_i, e_j) = 0 \) then \( \hat{\beta}_{OLS} \) is the most efficient estimator (Gauss-Markov Theorem)

Assumptions/requirements of OLS (2.3)

No relationship between explanatory variables and unobservables.

\( E[e_i | x_{2i}, x_{3i}, \ldots, x_{ki}] = 0 \Rightarrow \text{cov}(x_i, e_i) = 0 = \text{corr}(x_i, e_i) \).

Guarantees unbiasedness.

Variation within each explanatory variable and no multicollinearity.

\( \text{var}(x) \neq 0 \)

No perfect collinearity among variables

Necessary to calculate effects.

Constant variance and no cross-correlation (“ordinary error structure”)

\( \text{var}(e_i) = \sigma^2_e = \text{var}(e_j) \); \( \text{cov}(e_i, e_j) = 0 \)

Ensures efficiency

OLS 3 variation: Normality, \( e_i \sim N(0, \sigma^2_e) \).

Goodness of Fit

Total sum of squares: \( \sum (y_i - \bar{y})^2 = (N - 1) \cdot \text{var}(y) \)
Model sum of squares: \( \sum (\hat{y}_i - \bar{y})^2 = (N - 1) \cdot \text{var}(\hat{y}) \)

Residual sum of squares: \( \sum (\hat{y}_i - y_i)^2 = (N - 1) \cdot \text{var}(\hat{e}) \)

Coefficient of determination: \( R^2 = \frac{\text{MSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}} = \text{corr}(y, \hat{y})^2 \)

Interpretation
- Quality of model (useful, but incorrect)
- Comparison between models

Hypothesis testing (4.1, 4.2, 4.5)

Univariate: \( H_0 : \hat{\beta}_j = \beta^*_j; H_A : \hat{\beta}_j \neq \beta^*_j \)

Test statistic \( t^* = (\hat{\beta}_j - \beta^*_j) / \text{st.err.}(\hat{\beta}_j) \) has \( t \)-distribution with \( N - k \) d.o.f.

Multivariate
- Null hypothesis that a set of \( \beta \)s take on particular values;
- Alternative that at least one of them does not.
- Test statistic has \( F \)-distribution.
Example of Stata commands:

```
reg y x1 x2 x3           // OLS regression

test x1 = 2.13           // Test of individual hypothesis

test x2 = -5.9, accum    // Joint test of hypotheses

test x2 x3               // Joint test of equaling zero
```

```
. reg y x1 x2 x3

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5241.10184</td>
<td>3</td>
<td>1747.03395</td>
<td>F(  3,    96) = 1476.46</td>
</tr>
<tr>
<td>Residual</td>
<td>113.592501</td>
<td>96</td>
<td>1.18325521</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>5354.69434</td>
<td>99</td>
<td>54.0878216</td>
<td>R-squared = 0.9781</td>
</tr>
</tbody>
</table>

|                  | Coef.       | Std. Err. | t     | P>|t|     | [95% Conf. Interval] |
|------------------|-------------|------------|-------|---------|----------------------|
| x1               | 1.983209    | .109285    | 18.15 | 0.000   | 1.76628 - 2.200137   |
| x2               | -7.048816   | .1111005   | -63.45| 0.000   | -7.269349 - -6.828283|
| x3               | .0388324    | .107991    | 0.36  | 0.720   | -.1755282 - .2531929|
| _cons            | 3.109514    | .1091322   | 28.49 | 0.000   | 2.892888 - 3.32614   |

. test x1 = 2.13
( 1)  x1 = 2.13
    F(  1,    96) = 1.80
    Prob > F = 0.1824

. test x2 = -5.9, accum
( 1)  x1 = 2.13
( 2)  x2 = -5.9
    F(  2,    96) = 55.42
    Prob > F = 0.0000

. test x2 x3
( 1)  x2 = 0
( 2)  x3 = 0
    F(  2,    96) = 2076.70
    Prob > F = 0.0000
```

Top-left table: “SS” column contains the MSS, RSS, and TSS. Disregard “df” and “MS”.

Top-right table: number of observations; F-statistic for the “overall significance” of the regression (testing the hypothesis that all of the explanatory variables have zero effect); p-value of this hypothesis; $R^2$ of the regression. Disregard “adjusted $R^2$” and “Root MSE”.

Bottom table: “Coef.” column contains estimates of $\beta$; the next column has standard errors of each $\beta$; then the $t$-statistic testing the hypothesis that this variable has zero effect; then the p-value of this test; finally, a 95% confidence interval for the estimated coefficient.
Unit 2: Data Concerns

Collinearity (4.2)

Perfect collinearity: one explanatory variable is a linear function of others.

Implication: \( \hat{\beta} \) cannot be estimated.

Solution: Drop one variable; modify interpretation.

Near collinearity: high correlation between explanatory variables.

Implication: \( \hat{\beta} \) has large standard errors.

Solutions: Dropping variables (discouraged); Change nothing, but focus on joint significance (preferred).

Specification (6.2, 6.3)

Rescaling variables: no theoretical difference (some practical concerns) [6.1]

Omitted variables: Omitting \( x_3 \) from the model causes

\[
\mathbb{E}[\hat{\beta}_2] = \beta_2 + \frac{\text{cov}(x_2, x_3)}{\text{var}(x_2)} \beta_3 \quad \text{("omitted variable bias")}
\]

Irrelevant variables: Including irrelevant \( x_3 \) introduces no bias in estimation of \( \beta \), and \( \mathbb{E}[\hat{\beta}_3] = \beta_3 = 0 \).

Qualitative variables

Dummy variables: values of 0 or 1, depending on whether a condition is met.

Categorical variables: convert to a series of dummy variables; omit the "reference" category.

Nonlinear models

Common nonlinear specifications

Quadratics (for changing marginal effects)

\[
y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + e_i; \quad \Delta y/\Delta x = \beta_2 + \beta_3 x_i.
\]

Logarithms (for percentage changes and elasticities)

\[
y_i = \beta_1 + \beta_2 \ln(x_i) + e_i; \quad \beta_2 = \Delta y/\% \Delta x.
\]

\[
\ln(y_i) = \beta_1 + \beta_2 x_i + e_i; \quad \beta_2 = \% \Delta y/\Delta x.
\]

\[
\ln(y_i) = \beta_1 + \beta_2 \ln(x_i) + e_i; \quad \beta_2 = \% \Delta y/\% \Delta x.
\]

Interactions (for complementarities)

\[
y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 (x_{i2} \cdot x_{i3});
\]

\[
\Delta y/\Delta x_2 = \beta_2 + \beta_4 x_{i3} \quad \text{and} \quad \Delta y/\Delta x_3 = \beta_3 + \beta_4 x_{i2}.
\]
Interactions with dummy variables

Choosing a specification
- Economic theory (preferred)
- Eyeballing data
- Comparison of $R^2$ values (dangerous)

Testing a specification
- Simple: inclusion of higher order terms
- Ramsey’s Econometric Specification Error Test (RESET)

Dangers of “data mining” (and specification mining)

Classical measurement error (9.4)

True model: $y_i = \beta_1 + \beta_2 x_i + e_i$, but $\hat{x}_i = x_i + m_i$ measured instead.

“Classical”: $\mathbb{E}[m_i | \text{everything}] = 0$.

Implication: $\mathbb{E}[\hat{\beta}_{OLS}] = \beta \frac{\text{var}(x)}{\text{var}(x) + \text{var}(m)}$ (“attenuation bias”; “bias toward zero”)

Special case: tests of $H_0 : \beta_2 = 0$ unaffected.

Unusual observations (“outliers”) (9.5, 9.6)

Implication: OLS is highly sensitive to extreme values of $y_i$.

Solutions:
- Dropping outliers (dangerous)
- Least absolute deviations estimator: $\hat{\beta}_{LAD}$ to $\min \sum |y_i - x_i \hat{\beta}|$.
- No adjustments (recommended)

Interpretation of OLS results

Experimental data: researcher manipulates $x$ values.
- Correlation can be interpreted as causal effect.

Empirical data: generated through real-world processes
- Factors contributing to observed correlation, aside from effect of $x$ on $y$:
  - Unobserved heterogeneity
  - Reverse causality
  - Selection
Examples of Stata commands:

```stata
  gen d1 = (region == 1)                        Creation of dummy variable
  gen d2 = (region == 2)
  gen d3 = (region == 3)
  reg y x1 x2 x3 x4 d2 d3                      OLS with dummy variables
  test d2 d3                                    Joint test (“does region matter?”)

  gen xsquared = x^2                            Creation of squared term
  reg y x xsquared                              OLS estimation of quadratic

  gen logy=ln(y)                                Creation of log-variable
  gen logx=ln(x)
  reg logy logx                                 OLS estimation of elasticity

  gen inter12 = x1*x2                           Creation of interaction term
  reg y x1 x2 inter12                           OLS estimation with interaction

  qreg y x1 x2 x3                               LAD (“quantile” regression) estimation
```
Ramsey Econometric Specification Test

```
. reg y x1 x2 x3

Source |       SS       df       MS              Number of obs =  2134
----------|------------------------+----------------------------------+
Model | 35040.6408     3  11680.2136                           F(  3,  2130) = 437.98
Residual | 56803.8176  2130   26.668459                   Prob > F =  0.0000
----------|------------------------+----------------------------------+
Total | 91844.4584  2133  43.0588178                          R-squared =  0.3815
----------|------------------------+----------------------------------+

F(  3,  2130) = 437.98
Prob > F =  0.0000

y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------|------------------+------------------+-------+---------+------------------------+
x1 |  -1.185214    .1127196  -10.51   0.000     -1.406266    -0.9641618
x2 |   2.081589    .1122702   18.54   0.000     1.861418    2.301759
x3 |  -3.18763    .1100042  -28.98   0.000     -3.403357    -2.971904
_cons |  -0.528657    .1118070   -4.73   0.000     -0.7479189  -0.3093945
----------|------------------+------------------+-------+---------+------------------------+

Note: Despite the marginal significance of all the “yhat” terms in this example, they are jointly highly significant.
```