Econometrics - Cheat Sheet

Romain Fernex

February 2025

Contents

1

Regression model

1.1 Properties & Characteristics

1.1.1 What to know

- Adding a constant in a regression produces the same result as centering all variables first (including the DV) ! [1]
- $\hat{\theta}$ is what we call the OLS estimator. It is a statistic as it is a function of only observed variables.
- After standardization (simple regression) : the OLS estimator for $\hat{b}$ = correlation coefficient !
- multicollinearity : when one or more of the IVs are linear combination of the others
  - if this is true : X is singular (non invertible) so we can't solve for β_{OLS} !

Correlated vs Uncorrelated variables

- residuals($\hat{u}_i$) & regressors(x_i) : uncorrelated
  - proof : by construction
- residuals($\hat{u}_i$) & predicted values ($\hat{y}_i$) : uncorrelated
  - proof : take the covariance between the two and replace $\hat{y}_i$ by $\hat{a} + \hat{b}x_i$

1.1.2 Useful results :

1. $\bar{\hat{y}} = \bar{y}$
2. fitted/predicted values : $\hat{y}_i = \hat{a} + \hat{b}x_i$
3. sample variance : $Var_N(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$

[1] see pb 3 - pset 1

4. residuals : $\hat{u}_i = y_i - \hat{y}_i$
5. SST = SSE + SSR
6. $R^2 = \frac{SSE}{SST}$ (also called coefficient of determination)
   - $SST = \sum(y_i - \bar{y})^2$
   - $SSE = \sum(\hat{y}_i - \bar{y})^2$
   - $\sum(\hat{y}_i - y_i)^2$

1.1.3 The error terms

About error terms in a multiple regression

- Homoskedasticity and normality :

$$\begin{cases} \textbf{normality :}\ u_i \sim N(0, \sigma_0^2) \\ \text{by definition :}\ E(U|X) = 0 \iff \forall i, E(u_i|x_i) = 0 \\ \textbf{homoskedascity :}\ E(UU'|X) = I_n\sigma_0^2 \iff \forall i, E(u_i^2|x_i) = \sigma_0^2 \end{cases} \tag{1}$$

- Independent Identically Distributed (i.i.d)
- Important results :

$$\begin{cases} \textbf{OLS estimator :} \hat{\beta} \sim N(\beta_0, \sigma_0^2(X'X)^{-1}) \\ \textbf{fitted values :} \hat{y} \sim N(X\beta_0, \sigma_0^2 P_X) \\ \textbf{residuals :} \hat{u} \sim N(0_{Nx1}, \sigma_0^2 M_X) \end{cases} \tag{2}$$

- the mean square error (MSE) : ML estimator of σ_0^2 (to find it : just take the partial derivative of $L(\beta, \sigma^2)$ in σ)

$$\hat{\sigma}^2 = \frac{1}{N-K}\sum(y_i - x_i'\hat{\beta})^2 = \frac{SSR}{N-K} \text{ with K the number of parameters (including constant)} \tag{3}$$

1. $\hat{\sigma}$ is asymptotically unbiased (not necessarily only asymptotically for finite samples)
2. unbiased version : $\tilde{\sigma} = \frac{1}{N-K}\sum(y_i - x_i'\hat{\beta})^2$ and $\hat{\sigma}_{\hat{b}_k} = \sqrt{\tilde{\sigma}^2(X'X)^{-1}}$
3. $\frac{(N-K)\tilde{\sigma}}{\sigma_0^2} \sim \chi^2(N-K)$

Note that $E(U|X) = 0 \implies E(X) = 0$ but not reverse ! (proof using LIE)

1.2 Methods and results

1.2.1 A) Set up the minimization problem

The regular form of the problem is as follows : we note $\theta = (\beta_0...., \beta_n)$

$$\min_{\beta_0, \beta_1, .. \beta_n} SSR(\theta) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_n x_{in})^2 \tag{4}$$

We obtain the normal the normal equations by :

1. Taking the FOC of $SSR(\hat{\theta})$ with respect to each β_i
2. Substituting -2 by $\frac{1}{n}$ to get sample means

To ensure all variables have the same scale we can use standardization :

$$\forall i, \tilde{x}_i = \frac{x_i}{V_N(x_i)} \text{ and } \tilde{y}_i = \frac{y_i}{V_N(y_i)} \tag{5}$$

1.2.2 B) The matrix form : useful results

- key elements : for a system with K regressors and N observations

1. $x_i = \begin{pmatrix} 1 \\ x_{1i} \\ ... \\ x_{Ki} \end{pmatrix}$ and $X = \begin{pmatrix} x_1' \\ ... \\ x_n' \end{pmatrix}$
2. $\beta = \begin{pmatrix} a \\ b_1 \\ ... \\ b_K \end{pmatrix}$
3. Y' = y' = $\begin{pmatrix} y_1 & ... & y_n \end{pmatrix}$

- Key formulas :

1. Normal equation : $X'y = (X'X)\hat{\beta}$
   - $X'y = \sum x_i y_i$
   - $X'X = \sum x_i x_i'$
2. OLS estimator : $\hat{\beta} = (X'X)^{-1}X'y$
3. Predicted values and orthogonal projection matrix : $\hat{y} = X\hat{\beta} = P_X y$ with $P_X = X(X'X)^{-1}X'$

The M_X and P_X matrixes

Properties	**Associated formula**
– idempotent : $P_X^2 = P_X$	– $\hat{y} = P_X y$
– symmetric : $P_X^T = P_X$	– $\hat{u} = M_X y = M_X u$
– $M_X P_X = 0$	

1.2.3 C) Handling outliers

Two possible methods and their associated tradeoffs

1. log transformation : + reduces skewness in data distribution, linearize non linear relationships, helps with interpreting coefficients as elasticities / - careful with 0 values, sometimes makes interpretation complicated
2. trimming : + straightforward, preserves original scale of variable / - loss in information (decrease sample size), arbitrary

WARNING : Do not do both at the same time !!

1.2.4 D) Reading results

- Classic case :

1. $\beta_K = \frac{\Delta y}{\Delta x_K}$ (careful : this is a Delta, the partial derivative is only for continuous variables)
2. β_K is the marginal effect of x_K on y holding all other regressors constant.

3. **interpretation :** unit change in Y that results from a unit change in X[2]

- Log transformation case :

1. Principle : makes the estimation invariant to scale changes (changes of units). When multiplying x and y by α, the intercept is affected by α but not the slope coefficient
2. **Interpretation :**
   - Applied to IV only (ex : $Y = b_0 + b_1 ln(X) + U$) : A 1% change in X is associated with a change in Y of $0.01 * b_1$
   - Applied to DV only (ex : $ln(Y) = b_0 + b_1 X + U$) : A change in X by one unit ($\Delta X = 1$) is associated with a $(exp(b_1) - 1) * 100\%$ change in Y[3]
   - Applied to both IV and DV (ex : $ln(Y) = b_0 + b_1 ln(X) + U$) : A 1% change in X is associated with a $b_1\%$ change in Y, so b_1 is the elasticity of Y with respect to X.

- Non linear case (quadratic form : $a + b_1 x + b_2 x^2$..) :

1. Take the partial derivative of y in x to see whether the relation is decreasing or increasing as a function of x
2. Finding the x at which y is at its min/max based on this relation : just take the FOC of y wrt x

- Decomposition of variance : in case we scale a regressor (ex : $\tilde{x}_{1i} = x_{1i}/100$) then $\tilde{b}_1 std_N(\tilde{x}_{1i}) = b_1 std_N(x_{1i})$.

1. It means that while the coefficient itself changes when you change the unit of the regressor, its effect in terms of the overall spread (or variability) of the regressor remains the same.
2. In other words, the relative contribution of a regressor to your model doesn't depend on the specific units of measurement.

1.2.5 E) Handling interaction terms

- For dummy variables : Only keep the interaction terms that are of interest and retain standalone variables
- For continuous variable : Marginal effect of a regressor on the DV depends on the regressors with which we consider interaction terms

1.2.6 Ensuring that 2 IVs are uncorrelated :

- Dummy variable case : we ensure that we have the same share of each category of IV_1 in each category of IV_2

An intuitive example

- Every department (gender in our case) has exactly the same proportion of managers (diploma in our case)
- In this situation, knowing which department someone works in tells you nothing about their likelihood of being a manager, and vice versa
- This is exactly what zero covariance represents: no systematic relationship between the two characteristics

[2] be very careful about the scale/unit of both variables to properly interpret what this means
[3] the exponent transformation is not necessary if the coefficient is low enough

1.2.7 Disambiguation : On the difference between errors and residuals

1. Errors (u_i) :
   - Definition : The true, unobservable difference between the actual value and the true population regression line (which is the expression for the relation we are trying to estimate through the regression model). It represent inherent randomness in the true population.
     - why is there a difference between actual values and the true population parameter : actual (observed) values are sample dependent whereas the true population parameters do not change regardless of which sample we take.
     - it cannot be observed directly (we don't know the parameters from the true population !)
2. Residuals ($\hat{u}_i$) :
   - Definition : observable difference between the actual value and the predictor
     - formula : $\hat{u}_i = y_i - \hat{y}_i$
     - it can be estimated from sample data

What links boths ?

- Residuals are practical estimates of the unobservable error ! (residuals are proxies)
- As sample estimates approach the true parameters, the distribution of residuals approaches the distribution of errors.

2

Hypothesis testing

2.1 General methodology of hypothesis testing

1. Set up the null and alternative hypotheses
2. Set the test size
3. Give the decision rule :

Using the right decision rule

- **Two-tailed test**: Reject H_0 if |test statistic| > critical value
- **One-tailed test (right)**: Reject H_0 if test statistic > critical value
- **One-tailed test (left)**: Reject H_0 if test statistic < critical value
- **p-value approach**: Reject H_0 if p-value < significance level (α)

4. Compute the appropriate test statistic and determine its distribution based on : the size of the sample, what you know, what you are trying to test for
5. Compute the p-value / critical value
6. Compare the test statistic to the critical value

2.1.1 Special case : find the estimate of a coefficient in a constrained model

1. take $H_0 : b_1 = b2$ and $b_3 + b_1 = 1$ (alternative hypothesis : no constraints)
2. rewrite b_3, b_2 as a function of b_1
3. replace them in the regression equation and pull together the terms multiplied by b_1 (here $x_{1i}+x_{2i}-x_{3i}$
4. withdraw the extra terms from both sides (here x_{3i})
5. note X the terms multiplying b_1 and Y the expression on the left hand side
6. get the OLS estimator for a and b_1 and express those for b_2, b_3 as a function of $\hat{b}_1$

2.2 Important metrics

2.2.1 Level of significance (test size)

- Denotes the probability of a type I error (False positive)[4]
- Definition :

$$\alpha = \max_{\theta \in \Omega_0} P_\theta(\hat{W} > c) \tag{6}$$

2.2.2 Power of a test

- probability of rejecting H_0 for any value of $\theta \in \Omega_1$ (True positive)
- Definition :

$$\text{power} = P_{\theta \in \Omega_1}(\hat{W} > c_\alpha) \tag{7}$$

2.2.3 P-value

- simplified : probability that H_0 s true (if low enough one can reject H_0)

  - detailed : it tells you the probability of the data given the null hypothesis. In other words, the p-value is a conditional probability that assumes H_0 is true before seeing the data.
- alternative description : probability of observing a test statistic at least as extreme as the one calculated from your sample data, assuming the null hypothesis is true.
- Definition :

$$\text{p-value} : \max_{\theta \in \Omega_0} P_\theta(\hat{W} > \hat{W}(\omega)) \tag{8}$$

2.2.4 R-squared and adjusted R-squared

- what does it measures : share of the variation in the dependent variable that is explained by the independent variables considered in the model
- property : the regular R^2 goes up whenever you add a variable (does not depend on whether the variable in question is orthogonal to existing ones or not)
- formulas : with K the number of parameters (excluding the constant)

$$\begin{cases} R^2 = \dfrac{SSE}{SST} \\ \text{adjusted } R^2 = 1 - \dfrac{SSR/(N-K-1)}{SST/(N-1)} \end{cases} \tag{9}$$

[4]WARNING : it is not possible to minimize both Type I and Type II errors at the same time, there is always a tradeoff

2.2.5 Confidence intervals

- formula (for a t-test : $CI(\beta_0)_{1-\alpha} = [\hat{\beta} - t_\alpha \sigma_{\hat{\beta}} < \beta_0 < \hat{\beta} + t_\alpha \sigma_{\hat{\beta}}]$
- interpretation : the confidence interval contains the parameter β_0 with $(1-\alpha)\%$ confidence.

2.3 Useful distributions

2.3.1 Chi-square distribution

- condition : $X_1, ..X_n$ i.i.d with $X_i \sim N(\mu, \sigma^2)$
- property :

$$\sum_{i=1}^{n} (X_i - \bar{X})^2 / \sigma_2 \sim \chi^2(n-1) \tag{10}$$

- We have N-1 independent elements hence the N-1 degrees of freedom.

Proof

- We have the following constraint : $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$
- We can rewrite it to express $X_N - \bar{X}$ as a function of $\bar{X}$ and $(X_i)_{i=1}^{N-1}$

$$(X_N - \bar{X}) = (N-1)\bar{X} - \sum_{i=1}^{N-1} X_i \tag{11}$$

- properties : independence (if $X \sim \chi_k$ and $Y \sim \chi_n$ with X and Y independent $\implies X + Y \sim \chi_{k+n}$)

2.3.2 Student's t-distribution

- Condition : $X_1, ..X_n$ i.i.d and Formula : $\sim N(\mu, \sigma^2)$
- $\frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t(n-1)$
  - where s is the estimate for standard deviation
- Properties : symmetric around 0 + lower nb of df implies fatter tails (as nb of dg goes to ∞, this distribution converges to the normal distribution)

2.3.3 Fisher distribution

- Condition : $Q_1 \sim \chi^2(n_1)$ and $Q_2 \sim \chi^2(n_2)$, independent
- Formula : $\frac{Q_1/n_1}{Q_2/n_2} \sim F(n_1, n_2)$

2.4 Common tests

2.4.1 T-test

On the T-Test

What is it used for ? : test hypothesis on individual regression coefficients (works also with single linear restrictions such as "is $\beta_1 = \beta_2$")

- Set up : $H_0 : \beta_k = c$ with c a constant $[H_1 : \beta_k \neq c]$
- test statistic : $\hat{T} = \frac{\hat{b}_k}{\hat{\sigma}_{\hat{b}_k}} \sim t(n-k)$ with k the nb of parameters (including the constant)
- Decision rule : $|\hat{T}| > t_{1-\alpha/2}(n-k)$[a]

[a] $t_{1-\alpha}$ decreases in α

2.4.2 F-tests

On the Fisher Test

What is it used for ? : (1) test the joint significance[a] of several coefficients [for instance "is $\beta_1 = \beta_2 = 0$", we have 2 restrictions], (2) test the relevance of a given model

1. **Case 1 :** testing whether some parameters can be safely removed from the model[b]
   - Set up : H_0 : you can safely remove the m parameters of interest and take the restricted model without information loss $[H_1 : \exists k | \beta_k \neq 0]$
   - Test statistic : $\hat{F} = \frac{(SSR_r - SSR_{ur})/m}{SSR_{ur}/(n-k)} \sim F(m, n-k)$ with k the nb of parameters in the unrestricted model (including the constant)
   - Decision rule (reject H_0 if): $\hat{F} > F_{1-\alpha}(m, n-k)$
2. **Case 2 :** test whether all parameters but the constant are equal to zero
   - Goal : test whether a model as a whole has any explanatory power beyond just the mean
   - Set up : $H_0 : \forall i, \beta_i = 0$ and $a \neq 0$ $[H_1 : \exists k | \beta_k \neq 0]$
   - Test statistic : $\hat{F} = \frac{R^2_{ur}(n-k)}{(1-R^2_{ur})(k-1)} \sim F(k-1, n-k)$ with k the number of parameters (including the constant)
   - Decision rule : $\hat{F} > F_{1-\alpha(k-1,n-k)}$ or $(k-1)\hat{F} > \chi^2_{1-\alpha}(k-1)$

[a] joint significance of coefficients means that at least one of the coefficients tested is different from 0, H1 : $\exists \beta_k \neq 0$
[b] see : link with the formula for the F-test

2.4.3 Wald test

On the Wald Test

What is it used for ? : (1) simultaneously test for several parameters using matrices, (2) test for linear combinations of parameters

- Set up : $H_0 : R\beta_0 = c$, (R is a matrix of dimension MxK) $[H_1 : R\beta_9 \neq c]$
  - Example : $g(\beta_0) = 0$ with g a continuous function, $g(\beta_0)$ a Mx1 vector and β_0 a Kx1 vector
- Deriving the t-stat : use this theorem (with X a vector of p rv's and Σ non-sing.)

$$\text{if } X \xrightarrow{d} N(0,\Sigma) \Longrightarrow X^T\Sigma^{-1}X \xrightarrow{d} \chi^2_p \tag{12}$$

- Test statistic : if $R(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \hat{\sigma}^2 R(X'X)^{-1}R')$ then $(R\hat{\beta} - c) \xrightarrow{d} N(0, \hat{\sigma}^2 R(X'X)^{-1}R')^a$

$$\hat{W} = (R\hat{\beta} - c)'[\hat{\sigma}^2 R(X'X)^{-1}R']^{-1}(R\hat{\beta} - c) \xrightarrow{d} \chi^2_M \tag{13}$$

- Link with t-test stat: $\hat{W} = \hat{T^2}$ (two side student t test is a wald test)
- Decision rule : $|\hat{W}| > \chi^2_{1-\alpha}(M)$ with $\chi^2_{1-\alpha}(M)$ the $1-\alpha$ quantile of the chi-squared distribution with M degrees of freedom.
- General property of the Wald test : consistent

[a] WARNING : if we are using the delta method (as in the example above with g), J depends on $g(\beta_0)$ so we need to take into account the hypothesis we made and evaluate it in 0 when computing the test statistic !

2.4.4 Consistency of a test

- a test is consistent if : test power $= P_{\theta\in\Omega_1}(\hat{W} > c_\alpha) \xrightarrow{n\to\infty} 1$
- aka : probability of rejecting the null hypothesis when false tends to one

2.5 Other useful results

- standard error :
  1. $SE(\hat{\beta}_1) = \sqrt{V(\hat{\beta})} = \sqrt{\sigma^2(X'X)^{-1}}$ [5]
     - alternative formula : $SE(\hat{\beta}_1) = \sqrt{\frac{SSR}{N-K}(X'X)^{-1}}$
  2. $SE(\bar{x}) = \sqrt{V(\bar{x})} = \sqrt{\frac{\sigma^2}{n}}$

3

[5] we do not divide by $\sqrt{n}$ as the sample size n is already factored into this calculation through the X'X matrix !

Asymptotic theory

3.1 Modes of convergence

3.1.1 Sure convergence ($\xrightarrow{s}$)

- definition :
$$\forall w \in \Omega, \forall \epsilon > 0, \exists n_0(w, \epsilon), \forall n \geq n_0 \Longrightarrow |X_n(w) - X(w)| < \epsilon \tag{14}$$
- Pointwise ($\forall w \in \Omega$)
- If n_0 does not depend on w then we have uniform convergence a.k.a the sequence of functions converges to a limit function in such a way that the speed of convergence does not depend on the point in the domain.

3.1.2 Almost sure convergence $\xrightarrow{a.s}$

- Definition :
$$P(w : \lim_{n \to \infty)} X_n(w) = X(w)) = 1 \tag{15}$$
- Pointwise except for a set of P-probabilities 0 (pointwise convergence is not verified for a negligible set of outcomes)

3.1.3 Convergence in probabilities $\xrightarrow{P}$

- Definition :
$$\forall \epsilon > 0, P(w : |X_n(w) - X(w)| > \epsilon) \xrightarrow[n \to \infty]{} 0 \tag{16}$$
- Difference from the limit converges to 0 in probability
- Infinitely many n for which $X_n(w) \neq X(w)$

3.1.4 Convergence in quadratic mean ($\xrightarrow{L^2}$)

- Definition :
$$\lim_{n \to \infty} E(|X_n - X|^2) = 0 \tag{17}$$

3.1.5 Convergence in distribution ($\xrightarrow{d}$)

- Definition :
$$\forall x \in \mathbb{R}, F_n(x) \to F(x) \tag{18}$$
- Pointwise
- Implies $E(f(X_n)) \to E(f(X))$ i.i.f f is bounded and continuous !

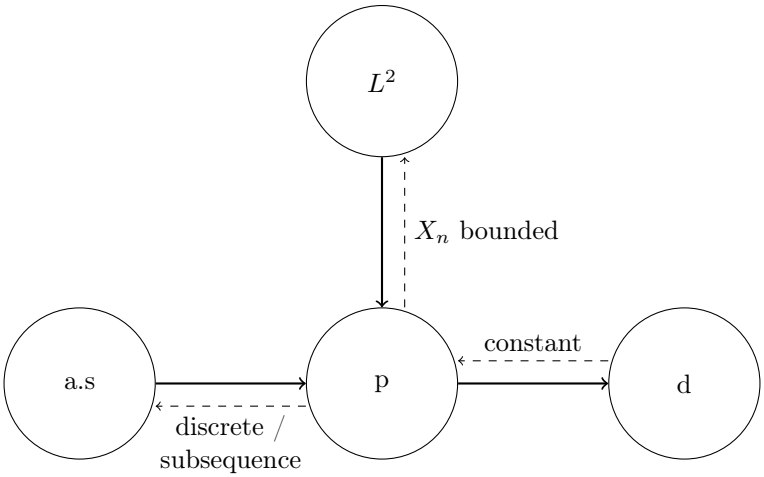

Figure 1: Relations between the different modes of convergence

3.2 Useful definitions and theorems

3.2.1 Consistency of an estimator

- **Definition :** Take $W_n = W_n(X_1, ...X_n)$ a sequence of estimators of θ, consistency implies :

$$\lim_{n\to\infty} P_\theta(|W_n - \theta| > \epsilon) = 0 \iff W_n \xrightarrow{P} \theta \tag{19}$$

- Conditions for existence :
  1. Asymptotically invariant : $\lim_{n\to\infty} V_\theta(W_n) = 0$
  2. Asymptotically unbiased : $\lim_{n\to\infty} E_\theta(W_n) = \theta \iff \lim_{n\to\infty} E_\theta(W_n - \theta) = 0$

Consistency of OLS (univ. / multiv.)

- univariate case :
  1. condition : i.i.d observations and $cov(u_i, x_i) = 0 (\iff E[u_i|x_i = 0])$ and $var(x_i) \neq 0$
  2. why ? :

$$\hat{b} = \frac{cov_n(y_i, x_i)}{var_n(x_i)} \xrightarrow{P} \beta_0 + \frac{cov(u_i, x_i)}{var(x_i)} \tag{20}$$

- multivariate case :
  1. condition : $E(x_i u_i) = 0 (\iff E[u_i|x_i = 0])$ and $E(x_i x_i')$ non singular
  2. why ? :

$$\hat{\beta} = \beta_0 + \frac{1}{N}(\sum_{i=1}^{N} x_i x_i')^{-1} \frac{1}{N}(\sum_{i=1}^{N} x_i u_i) \xrightarrow{P} \beta_0 + E(x_i x_i')^{-1} E(x_i u_i) \tag{21}$$

3.2.2 Slutsky theorem (P,d)

- if $X_n \xrightarrow{d} X, Y_n \xrightarrow{P} c \implies X_n + Y_n \xrightarrow{d} X + c$ and $X_n Y_n \xrightarrow{d} cX$[6]

[6]The weaker form of convergence dominates (here d)

3.2.3 Continuous mapping theorem (a.s, P, d)

- if $X_n \to X$ then $g(X_n) \to g(X)$
- Condition : g almost continuous (at least)
- concrete example : given $E(x_i x_i')$ non singular, $f(A) = A^{-1}$ is continuous on the set of non singular matrix so : $(\frac{1}{N}\sum_{i=1}^{N} x_i x_i')^{-1} \xrightarrow{P} E(x_i x_i')^{-1}$

3.2.4 Asymptotic equivalence (P $\to d$)

- if $|X_n - Y_n| \xrightarrow{P} 0$ and $X_n \xrightarrow{d} X$ then $Y_n \xrightarrow{d} X$

3.2.5 Law of large numbers (P, a.s)

- Condition : X_i i.i.d + mean is finite (for strong law $E(|X_i|$ must be finite)
- Weak/strong law : $\frac{1}{N}\sum_{i=1}^{N} X_i \xrightarrow{P/a.s} E(X_i) = \mu$

3.2.6 Central Limit Theorem (d)

- Condition : X_is i.i.d (here, mean is μ and covariance matrix is Σ)
- Expression : $\sqrt{N}(\bar{X} - \mu) \xrightarrow{d} N(0, \Sigma)$

3.3 Methods and results

3.3.1 Applying the Delta method(d)

- Take a_n a sequence depending on n, X_n a sequence of rvs and m a constant.
- Property :

$$a_n(X_n - m) \xrightarrow{d} Z \Longrightarrow a_n(f(X_n) - f(m)) \xrightarrow{d} J(m)Z \text{ with } J(m) = (\frac{\delta f_i(m)}{\delta x_j})_{i,j} \tag{22}$$

3.3.2 Useful results

- Sample mean : $\bar{X} = \frac{1}{N}\sum X_i$ and $\bar{X} \xrightarrow{P} N(\mu, \theta)$
- Sample variance : $V(\bar{X}) = \frac{1}{N}\sum V(X_i)$ with $V(X_i) = \sigma^2$ if X_n are i.i.d
- estimated variance of $\hat{\beta}$: $V(\hat{\beta}) = \sqrt{\tilde{\sigma}^2(X'X)^{-1}} = \sqrt{\frac{1}{N-K}\sum(y_i - x_i'\hat{\beta})^2}$

Common Asymptotic Distributions

- OLS(univ.) : $\sqrt{N}(\hat{b} - b_0) \xrightarrow{d} N(0, \frac{\sigma_0^2}{Var(x_i)})$
- OLS (multiv.) : $\sqrt{N}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, \sigma_0^2 E(x_i x_i')^{-1}) = N(0, \sigma_0^2 (X'X)^{-1})$
- t-test : $t(n) \xrightarrow{d} N(0, 1)$
- F-test : $kF(k, n) \xrightarrow{d} \chi_k^2$ (so $k\hat{F}$ is asymptotically equivalent to the wald statistic)

4

Appendix

4.1 Computing estimates from the Covariance matrix

Multiple vs Simple regression Case

1. Simple regression case : (model of the form : $y_i = \alpha + \beta x_i + u_i$)

   - **Coefficient estimate** :
   $$\hat{\beta} = \frac{cov(x,y)}{var(x)} \tag{23}$$

   - **Standard error of the coefficient :** [a]
   $$\begin{aligned} SE(\hat{\beta}) &= \sqrt{\hat{\sigma}^2 (X'X)^{-1}} \\ &= \sqrt{\frac{SSR}{N-K} var(x)^{-1}} \\ &= \sqrt{\frac{var(y) - \hat{\beta}^2 var(x)}{(N-K)var(x)}} = \sqrt{\frac{var(y) - \hat{\beta} cov(x,y)}{(N-K)var(x)}} \end{aligned} \tag{24}$$

2. Multivariate regression case : (model of the form : $y_i = \beta x_i' + u_i$)

   - **Coefficient estimate :** we note Σ_{XX} the covariance matrix between IVs and ΣXY the covariance matrix of the IVs with the DV[b]
   - simple version (two regressors X,Z only)[c] :
   $$\hat{\beta_x} = \frac{var(z)cov(x,y) - cov(x,z)cov(z,y)}{var(x)var(z) - cov(x,z)^2} \tag{25}$$

   - general version (more than 2 regressors) : [d]
   $$\hat{\beta_j} = \frac{det(\Sigma_{XX}^{(j)})}{det(\Sigma_{XX})} \text{ where } \Sigma_{XX}^{(j)} \text{ s the matrix obtained by replacing the } j^{th} \text{ column of } \Sigma_{XX} \text{ by } \Sigma_{XY} \tag{26}$$

PS : We cannot compute the SE from the covariance matrix of variables in the multivariate case !

[a] for full derivation see my extended notes + problem set 2 exercise 3

[b] $\Sigma_{XY} = \begin{pmatrix} cov(x_1, y) \\ .. \\ cov(x_n, y) \end{pmatrix}$

[c] for the intuition see Pset 1 exercise 4

[d] look up Cramer's rule to understand why that works ! For the full derivation see Basu Deepankar, *The Yule-Frisch-Waugh-Lovell Theorem for Linear Instrumental Variables Estimation*, p38-39

4.2 A bit of matrix algebra

- For variance computations :
  1. $V(BU) = BV(U)B'$ with B a deterministic matrix and U a vector of random variables
  2. $V(X) = E(XX') - E(X)E(X') = E[(X - EX)(X - EX)']$
  3. using the LIE : $V(X) = E(E((X - EX)^2|X))$
- Playing with transposition :

1. $E(X') = E(X)'$
2. $(A+B)' = A' + B'$
3. $(A^{-1})' = (A')^{-1}$
4. $(AB)' = B'A'$
5. For any matrix A : A'A is positive semidefinite ($\Longrightarrow A'A \geq 0$)

Proof

Proof of Positive Semidefiniteness:
For any vector $x \in \mathbb{R}^m$,with $M = AA^T$

$$x^T M x = x^T (AA^T)x = (A^T x)^T (A^T x) = \|A^T x\|^2 \geq 0.$$

Since the squared norm is always non-negative, M is positive semidefinite.

- Homoskedacity in matrix form : $V(U|X) = \sigma^2 I$

4.3 Other useful formulas

- Unbiased variance estimator under homoskedacity : $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-K}$ with K the number of regressors (including the intercept!)

4.4 Table with common distributions

Table 1: Probability Distributions – Part I

Distribution	Type	Support	PDF/PMF
Uniform (Cont.) $\mathcal{U}(a,b)$	Cont.	$x \in [a,b]$	$f(x) = \frac{1}{b-a}$
Uniform (Disc.) $\mathcal{U}\{a,\dots,b\}$	Disc.	$x \in \{a,\dots,b\}$	$P(X=x) = \frac{1}{b-a+1}$
Normal $N(\mu,\sigma^2)$	Cont.	$x \in \mathbb{R}$	$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
Exponential $\text{Exp}(\lambda)$	Cont.	$x \in [0,\infty)$	$f(x) = \lambda e^{-\lambda x}$
Poisson $\text{Pois}(\lambda)$	Disc.	$x \in \{0,1,2,\dots\}$	$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$
Geometric $\text{Geom}(p)$	Disc.	$x \in \{1,2,3,\dots\}$	$P(X=x) = (1-p)^{x-1}p$
Binomial $\text{Bin}(n,p)$	Disc.	$x \in \{0,1,\dots,n\}$	$P(X=x) = \binom{n}{x}p^x(1-p)^{n-x}$
Hypergeometric $\text{Hyp}(N,K,n)$	Disc.	$x \in \{\max(0,n-(N-K)),\dots,\min(n,K)\}$	$P(X=x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$
Bernoulli $\text{Bern}(p)$	Disc.	$x \in \{0,1\}$	$P(X=x) = p^x(1-p)^{1-x}$

Table 2: Probability Distributions – Part II

Distribution	**CDF**	**Expectation**	**Variance**
Uniform (Cont.) $\mathcal{U}(a,b)$	$F(x) = \frac{x-a}{b-a}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Uniform (Disc.) $\mathcal{U}\{a,\dots,b\}$	$F(x) = \frac{\lfloor x \rfloor - a + 1}{b-a+1}$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$
Normal $N(\mu,\sigma^2)$	$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$	μ	σ^2
Exponential $\text{Exp}(\lambda)$	$F(x) = 1 - e^{-\lambda x}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$
Poisson $\text{Pois}(\lambda)$	$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \frac{\lambda^k e^{-\lambda}}{k!}$	λ	λ
Geometric $\text{Geom}(p)$	$F(x) = 1 - (1-p)^{\lfloor x \rfloor}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
Binomial $\text{Bin}(n,p)$	$F(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$	np	$np(1-p)$
Hypergeometric $\text{Hyp}(N,K,n)$	$F(x) = \sum_{k=\max(0,n-(N-K))}^{\lfloor x \rfloor} \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$	$\frac{nK}{N}$	$\frac{nK(N-K)(N-n)}{N^2(N-1)}$
Bernoulli $\text{Bern}(p)$	$F(x) = \begin{cases} 0, & x < 0 \\ 1-p, & 0 \le x < 1 \\ 1, & x \ge 1 \end{cases}$	p	$p(1-p)$

4.5 Table of common estimators

Table 3: Summary of Statistical Estimators and Their Distributions: Part I

Known/Assumed Information	**Parameter/Hypothesis**	**Test Statistic**
σ^2 known	Mean μ	$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$
μ known	Variance σ^2	$\chi^2 = \frac{n\sum(x_i - \mu)^2}{\sigma^2}$ (or, for a sample, $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$)
Neither known	Mean μ	$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$
Neither known	Variance σ^2	$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$
σ^2 known	Proportion p	$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$
Neither known	Difference in Means ($\mu_1 - \mu_2$)	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Multivariate case; Covariance matrix known	Vector of Means μ	$T^2 = n(\bar{x} - \mu)'\Sigma^{-1}(\bar{x} - \mu)$
Regression (parameters unknown)	Individual Coefficient β	$t = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$
Regression (overall test)	Overall Model Significance (e.g. $H_0 : \beta_1 = \beta_2 = \cdots = 0$)	$F = \frac{MS_{\text{regression}}}{MS_{\text{error}}}$

Table 4: Summary of Statistical Estimators and Their Distributions: Part II

Known/Assumed Information	**Distribution**
σ^2 known	$N(0,1)$
μ known	$\chi^2_{(n)}$ or $\chi^2_{(n-1)}$
Neither known	$t_{(n-1)}$
Neither known	$\chi^2_{(n-1)}$
σ^2 known	$N(0,1)$
Neither known	$t_{(n_1+n_2-2)}$
Multivariate case; Covariance matrix known	$\chi^2_{(p)}$
Regression (parameters unknown)	$t_{(n-k)}$
Regression (overall test)	$F_{(k-1,n-k)}$

4.6 Hypothesis Testing Reference Tables

Test Type	**Null Hypothesis**	**Test Statistic**	**Distribution**	**When to Use**
One-sample z-test	$\mu = \mu_0$	$W = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$	Standard Normal (Z)	Testing a population mean with known variance
One-sample t-test	$\mu = \mu_0$	$\hat{T} = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$	t-distribution with $(n-1)$ df	Testing a population mean with unknown variance
Two-sample t-test (independent)	$\mu_1 = \mu_2$	$t = \frac{\bar{x}_1-\bar{x}_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}$	t-distribution with df calculated using Welch's approximation	Comparing means of two independent groups
One-proportion z-test	$p = p_0$	$z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Standard Normal (Z)	Testing a population proportion
Two-proportion z-test	$p_1 = p_2$	$z = \frac{\hat{p}_1-\hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$	Standard Normal (Z)	Comparing proportions from two populations
F-test (variance ratio)	$\sigma_1^2 = \sigma_2^2$	$\hat{F} = \frac{s_1^2}{s_2^2}$	F-distribution with (n_1-1, n_2-1) df	Comparing variances of two populations

Table 5: Common Hypothesis Tests

Test Type	**Null Hypothesis**	**Test Statistic**	**Distribution**	**When to Use**
t-test for regression coefficient	$\beta = c$ (a constant)	$t = \frac{\hat{\beta}-c}{SE(\hat{\beta})} = \frac{\hat{b_k}}{\sigma \hat{b}_k}$ [7]	t-distribution with $(n-k)$ df	Testing significance of individual regression coefficients
F-test for relevance of overall regression model	All $\beta = 0$ (excluding constant)	$\hat{F} = \frac{(TSS-RSS)(n-k)}{(1-RSS)(k-1)}$ [8]	F-distribution with $(k-1, n-k)$ df	Testing if any predictors are significant
F-Test (restricted vs unrestricted model)	β_1 and β_2 are equal to 0	$\hat{F} = \frac{(R^2_{ur}-R^2_r)/q}{(1-R^2_{ur})/(n-k)}$ [9]	F-distribution with $(q, n-k)$ df	Testing for joint significance of two variables

Table 6: Regression-Based Tests

[7] see : note about error terms page 2 for more details on the SE

[8] see : page 90 in this book

[9] q is the number of restrictions, k is the number of predictors in the unrestricted model(including constant)