Základy ekonometrie

IV. Model vícenásobné regrese

Obsah tématu

Úvod

- Prohloubení znalostí o modelu vícenásobné regrese.
- Maticové vyjádření problému.
- Ilustrativní důkazy.
- Rozšíření metod pro testování hypotéz.

Obsah tématu

Klasické předpoklady

1. $E(\epsilon_i) = 0$. Nulová střední hodnota náhodných složek.
2. $var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$. Konstantní rozptyl náhodných složek (homoskedasticita).
3. $cov(\epsilon_i, \epsilon_j) = 0$ pro $i \neq j$. ϵ_i a ϵ_j jsou vzájemně nekorelované.
4. ϵ_i má normální rozdělení.
5. $X_{1i}, \ldots, X_{ki}$ jsou pevně daná, jedná se o nenáhodné veličiny.

LRM v maticovém vyjádření I

- k vysvětlujících proměnných $x_{i1}, \ldots, x_{ik}$ pro $i = 1, \ldots, N$ a model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

- Úrovňová konstanta: $x_{i0} = 1$.
- Vektory $N \times 1$ a $k + 1 \times 1$:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_N \end{bmatrix} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \cdot \\ \cdot \\ \epsilon_N \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \beta_k \end{bmatrix}$$

LRM v maticovém vyjádření II

- Matice vysvětlujících proměnných rozměru $N \times k + 1$

$$X = \begin{bmatrix} 1 & x_{11} & \cdot & \cdot & x_{1k} \\ 1 & x_{21} & \cdot & \cdot & x_{2k} \\ \cdot & \cdot & \cdot\cdot & \cdot & \\ \cdot & \cdot & \cdot\cdot & \cdot & \\ 1 & x_{N1} & \cdot & \cdot & x_{Nk} \end{bmatrix}$$

- Lineární regresní model:

$$y = X\beta + \epsilon.$$

Odhad parametrů – dvě vysvětlující proměnné

- Model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

- Minimalizace součtu čtverců reziduí:

$$\widehat{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2},$$

$$\widehat{\beta}_2 = \frac{\left(\sum x_{2i} y_i\right)\left(\sum x_{1i}^2\right) - \left(\sum x_{1i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2},$$

$$\widehat{\alpha} = \overline{Y} - \widehat{\beta}_1 \overline{X}_1 - \widehat{\beta}_2 \overline{X}_2,$$

kde

$$y_i = Y_i - \overline{Y},$$

$$x_{1i} = X_{1i} - \overline{X}_1,$$

$$x_{2i} = X_{2i} - \overline{X}_2.$$

OLS estimátor – maticové vyjádření

- Minimalizace $SSR = (y - X\widehat{\beta})'(y - X\widehat{\beta}) = \widehat{\epsilon}'\widehat{\epsilon}$:

$$\widehat{\beta} = (X'X)^{-1}X'y$$

- Vlastnosti: lineární, nestranný, vydatný $\Rightarrow$ BLUE (Gaussův-Markovův teorém).

OLS odhad rozptylu náhodných složek

- Nestranný estimátor pro rozptyl náhodných složek, σ^2:

$$s^2 = \frac{\sum \widehat{\epsilon}_i^2}{N - k - 1},$$

kde

$$\widehat{\epsilon}_i = Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \ldots - \widehat{\beta}_k X_{ki}$$

jsou OLS rezidua.

OLS odhad rozptylu odhadů parametrů – dva regresory

- V případě $k = 2$ je rozptyl OLS odhadů:

$$var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{(1 - r^2)\sum x_{1i}^2},$$

$$var\left(\widehat{\beta}_2\right) = \frac{\sigma^2}{(1 - r^2)\sum x_{2i}^2},$$

  kde r je (výběrový) koeficient korelace mezi X_1 a X_2.
- V praxi nahrazujeme σ^2 příslušným odhadem, s^2.
- Využití při testování hypotéz.

OLS odhad rozptylu odhadů parametrů – obecně

- Kovarianční matice odhadu *vektoru* parametrů, $\widehat{\beta}$:

$$var(\widehat{\beta}) = \sigma^2 (X'X)^{-1}.$$

- Rozptyl náhodných složek, σ^2, nahrazujeme v praxi OLS odhadem.
- Rozptyly jednotlivých odhadů parametrů – prvky na diagonále kovarianční matice.
- Důkaz (při splnění klasických předpokladů):

$$\begin{aligned} var\left(\widehat{\beta}\right) &= E\left[\left(\widehat{\beta}-\beta\right)\left(\widehat{\beta}-\beta\right)'\right] = E\left[\left(X'X\right)^{-1}X'\epsilon\epsilon'X\left(X'X\right)^{-1}\right] \\ &= \left(X'X\right)^{-1}X'E\left[\epsilon\epsilon'\right]X\left(X'X\right)^{-1} \\ &= \left(X'X\right)^{-1}X'\left(\sigma^2 I_N\right)X\left(X'X\right)^{-1} = \sigma^2\left(X'X\right)^{-1}. \end{aligned}$$

Test významnosti při známém σ^2

1. Specifikace nulové hypotézy H_0 a alternativní hypotézy H_1.
2. Specifikace testové statistiky.
3. Specifikace rozdělení testové statistiky za předpokladu platnosti nulové hypotézy.
4. Volba hladiny významnosti.
5. Využitím kroků 3 a 4 získáme kritickou hodnotu.
6. Výpočet testové statistiky z kroku 2 a její porovnání s kritickou hodnotou z kroku 5. H_0 zamítáme v případě, kdy je absolutní hodnota testové statistiky větší než kritická hodnota (v opačném případě nezamítáme).

Test významnosti při známém σ^2 – příklad

1. Regrese se dvěma vysvětlujícími proměnnými; $H_0 : \beta_2 = 0$ a $H_1 : \beta_2 \neq 0$.
2. Pro hypotézu z kroku 1 je obvyklou testovou statistikou
$$Z = \frac{\widehat{\beta}_2 - \beta_2}{\sqrt{var\left(\widehat{\beta}_2\right)}} = \frac{\widehat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}}}.$$
3. Analogická odvození z dřívějška:
$$Z = \frac{\widehat{\beta}_2}{\sqrt{\frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}}} \sim N(0,1).$$
4. Provedeme obvyklou volbu 5 % (0.05).
5. Z odpovídá $N(0,1)$ a $\Pr[-1.96 \leq Z \leq 1.96] = 0.95 \rightarrow$ kritická hodnota 1.96.
6. V našem příkladu zamítáme H_0 pokud $|Z| > 1.96$.

t-test

- Neznámé σ^2 nahrazujeme odhadem (OLS rozptylem reziduí).
- Testová statistika má Studentovo t-rozdělení s $N - k - 1$ stupni volnosti (počet pozorování mínus počet odhadovaných parametrů).
- Pro regresi se dvěma vysvětlujícími proměnnými, $H_0 : \beta_2 = 0$:

$$t = \frac{\widehat{\beta}_2}{\sqrt{\frac{s^2}{(1-r^2)\sum x_{2i}^2}}} \sim t_{N-k-1}.$$

- Výhodné využití p-hodnot.

Koeficient determinace

- Měřítko kvality modelu $\rightarrow$ soulad modelu s daty.

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum \widehat{\epsilon}_i^2}{\sum \left(Y_i - \overline{Y}\right)^2}.$$

- Interpretace: podil variability závisle proměnné, která je vysvětlena (variabilitou či chováním) vysvětlujících proměnných.
- Interpretace jen v případě přítomnosti úrovňové konstanty!!! (jinak $TSS \neq RSS + SSR$)
- S přidáním další vysvětlující proměnné nikdy neklesne.

Korigovaný koeficient determinace

- Korigovaný koeficient determinace, $\overline{R}^2$:

$$\overline{R}^2 = 1 - \frac{\frac{SSR}{N-k-1}}{\frac{TSS}{N-1}} = 1 - \frac{n-1}{n-k-1}\left(1-R^2\right) = 1 - \frac{s^2}{\frac{1}{N-1}\sum\left(Y_i - \overline{Y}\right)^2}.$$

- Zohlednění přidání nevýznamných proměnných.
- Podobná motivace jako R^2 $\times$ nelze interpretovat tak, že odpovídá podílu variability závisle proměnné, kterou lze vysvětlit chováním vysvětlujících proměnných.
- Ve vztahu je podíl rozptylu náhodných složek a výběrového rozptylu závisle proměnné.
- Vždy menší nebo roven R^2; s přidáním málo významné proměnné může klesnout $\times$ s přidáním významné vysvětlující proměnné může i vzrůst.

Obsah tématu

Omitted variable bias – skutečnost

- Skutečný model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

- Korektní OLS odhad:

$$\widehat{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}.$$

- Malá písmenka – odpovídající odchylky od průměrů.

Omitted variable bias – opomenutí

- Model, se kterým pracujeme:

$$Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i.$$

- Odhad parametru β_1:

$$\tilde{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2},$$

- $\tilde{\beta}_1$ je vychýlený estimátor (není již tedy nestranný).

Omitted variable bias – důkaz

- Lze ukázat:

$$E\left(\tilde{\beta}_1\right) = E\left(\beta_1 + \frac{\beta_2 \sum x_{1i}x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i}\left(\epsilon_i - \bar{\epsilon}\right)}{\sum x_{1i}^2}\right)$$
$$= \beta_1 + \frac{\beta_2 \sum x_{1i}x_{2i}}{\sum x_{1i}^2}.$$

- $\tilde{\beta}_1$ je vychýlený estimátor (není již tedy nestranný).

Omitted variable bias – komentář

- Z předchozího výrazu: zkreslení nenastává v případě, kdy je $\beta_2 = 0$ nebo $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2}$.
- První případ nezajímavý (pokud je $\beta_2 = 0$, potom X_2 není ve skutečné regresi a nedošlo k opomenutí).
- Výraz $\frac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2}$ je úzce spojen s korelací mezi X_1 a X_2, kterou označíme jako r.
- Zkreslení při nezahrnutí důležité proměnné nenastává v případě, pokud je nezahrnutá vysvětlující proměnná nekorelována se zahrnutou vysvětlující proměnnou.
- Při znalosti problematiky možno vyslovit soudy o směru zkreslení (např. cena domu a atraktivity oblasti).

Zahrnutí nepodstatné proměnné

- Skutečný model: $Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i$.
- Chybná specifikace: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$.
- Chybný estimátor:

$$\tilde{\beta}_1 = \frac{\left(\sum x_{1i} y_i\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{2i} y_i\right)\left(\sum x_{1i} \sum x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}.$$

- Korektní estimátor:

$$\widehat{\beta}_1 = \frac{\sum x_{1i} y_i}{\sum x_{1i}^2}.$$

- Pokud ukážeme nestrannost $\tilde{\beta}_1$, lze s odkazem na Gaussův-Markovův teorém říct, že $var\left(\tilde{\beta}_1\right) > var\left(\widehat{\beta}_1\right)$.
- Zahrnutí irelevantní vysvětlující proměnné vede k méně přesným odhadům.

Multikolinearita – úvod

- Vysvětlující proměnné navzájem silně korelovány $\Rightarrow$ nesou v sobě zhruba tutéž informaci $\Rightarrow$ OLS estimátor má problém v odhadu oddělených mezních vlivů pro takto silně korelované proměnné.
- Nepřesný odhad koeficientů i v případě, kdy vysvětlující proměnné mohou mít společně velkou vysvětlující sílu.
- Obvyklým řešením: vypuštění jedné z vysoce korelovaných vysvětlujících proměnných.

Multikolinearita – detaily

- Rozptyly OLS estimátorů v modelu vícenásobné regrese se dvěma vysvětlujícími proměnnými:

$$var\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{(1-r^2)\sum x_{1i}^2},$$
$$var\left(\widehat{\beta}_2\right) = \frac{\sigma^2}{(1-r^2)\sum x_{2i}^2}.$$

- Vztahy vstupují do odvození intervalů spolehlivosti a do postupů testování hypotéz.
- Vystupuje zde korelační koeficient, $r \Rightarrow$ pokud perfektní multikolinearita ($r = 1$ nebo $r = -1$), rozptyly nejsme schopni vypočítat (a stejně tak i odhady).
- Dokonalá multikolinearita: matice $(X'X)$ singulární $\rightarrow$ neexistuje inverze.

Přibližná multikolinearita

- Výraz $(1 - r^2)$ blízký nule.
- Obecně kovarianční matice $\sigma^2(X'X)^{-1}$ „velká“ $\rightarrow$ vysoké směrodatné odchylky $\Rightarrow$ nepřesné odhady rozptylů odhadů parametrů.
- Malé t-statistiky, široké intervaly spolehlivosti.
- Nemá vliv na koeficient determinace $\rightarrow$ regrese dobře vystihne chování dat (někdy všechny parametry statisticky nevýznamné + vysoký koeficient determinace).

„Testování“ multikolinearity

- Pro rozptyl odhadu parametru lze ukázat (viz např. Heij et al. (2004), str. 157-159):

$$var\left(\widehat{\beta}_j\right) = \frac{\sigma^2}{(n-1)s_{x_j}^2\left(1-R_j^2\right)},$$

- pro $j = 2, \ldots, k$, kde R_j^2 je R^2 pomocné regrese j-tého regresoru na zbylých $(k-1)$ regresorů (vč. konstanty), $s_{x_j}^2$ je výběrový rozptyl x_j.
- Faktor zvyšující rozptyl – variance inflation factor (VIF): $\frac{1}{1-R_j^2}$.
- Hodnoty VIF větší než 10 mohou indikovat problém.

Obsah tématu

Úvod

- Obecný model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$

- Test hypotéz zahrnující více parametrů (jejich kombinaci).
- F-testy a testy založené na věrohodnostním poměru (širší uplatnitelnost).

Obsah tématu

Základní testovaná hypotéza

- Test $R^2 = 0$ ekvivalentní testu hypotézy:

$$H_0 : \beta_1 = \ldots = \beta_k = 0.$$

- není totožná s testováním k samostatných hypotéz $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ až $H_0 : \beta_k = 0$.
- F-statistika pro model vícenásobné regrese s k vysvětlujícími proměnnými a úrovňovou konstantou:

$$F = \frac{R^2}{1 - R^2} \frac{N - k - 1}{k}.$$

- Při platnosti nulové hypotézy má F-statistika rozdělení $F_{k,N-k-1}$.
- Vyhodnocení testu: kritické hodnoty testové statistiky, p-hodnoty testu.

Rozšíření testů – úvod

- Příklad modelu vícenásobné regrese se třemi vysvětlujícími proměnnými.
- Původní regresní modelu: *neomezený model (unrestricted model)*.
- Regresní model se zahrnutím restrikcí vyplývajících z formulované hypotézy: *omezený model (restricted model)*.
- Neomezený model:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \epsilon_i.$$

- Příklad hypotézy:

$$H_0 : \beta_1 = \beta_2 = 0.$$

- Jakoukoliv lineární funkci regresních koeficientů: $a\beta_1 + b\beta_2 + c\beta_3 = d$ pro nějaké konstanty a, b, c a d.
- Výsledný omezený model:

$$Y_i = \alpha + \beta_3 X_{3i} + \epsilon_i.$$

Rozšíření testů – příklady

- Obecnější hypotézy:

$$H_0 : \beta_1 = 0, \quad \beta_2 + \beta_3 = 1.$$

- Druhé z omezení lze zapsat jako $\beta_2 = 1 - \beta_3$.
- Omezený model:

$$Y_i - X_{2i} = \alpha + \beta_3 (X_{3i} - X_{2i}) + \epsilon_i.$$

- Odpovídá jednoduchému regresnímu modelu se závisle proměnnou $Y - X_2$, úrovňovou konstantou a vysvětlující proměnnou $(X_{3i} - X_{2i})$.

Rozšíření testů – obecně

- Možno implementovat lineární omezení do nového modelu (jiné proměnné).
- Testová statistika:

$$F = \frac{(SSR_R - SSR_{UR})/q}{SSR_{UR}/(N-k-1)}.$$

- SSR je součet čtverců reziduí, dolní indexy UR (neomezený model) a R (omezený model).
- Počet testovaných omezení je q.
- Intuice: „velké“ hodnoty F naznačují, že H_0 není korektní.
- F má Fischerovo-Snedecerovo rozdělení, $F_{q,N-k-1}$.
- F-statistika pomoci koeficientů determinace (jen pro stejné závisle proměnné):

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1-R^2_{UR})/(N-k-1)}.$$

Obsah tématu

Obsah tématu

Nelinearita v regresi

- LRM:
$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i.$$
- Nelineární model:
$$Y_i = f\left(X_{1i}, \ldots, X_{ki}, \alpha, \beta_1, \ldots, \beta_k\right) + \epsilon_i,$$
kde $f(\cdot)$ je nějaká nelineární funkce vysvětlujících proměnných a parametrů.
- Odlišná interpretace parametrů než u LRM.
- Odhad metodou maximální věrohodnosti:
$$L\left(\alpha, \beta_1, \ldots, \beta_k\right) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - f\left(X_{1i}, \ldots, \beta_k\right)\right)^2\right].$$
- Obecně nejsou estimátory v algebraické podobě! (numerická optimalizace).

Transformace modelu

- Cobb-Douglasova produkční funkce:

$$Y_i = \alpha_1 X_{1i}^{\beta_1} X_{2i}^{\beta_2} \dots X_{ki}^{\beta_k}.$$

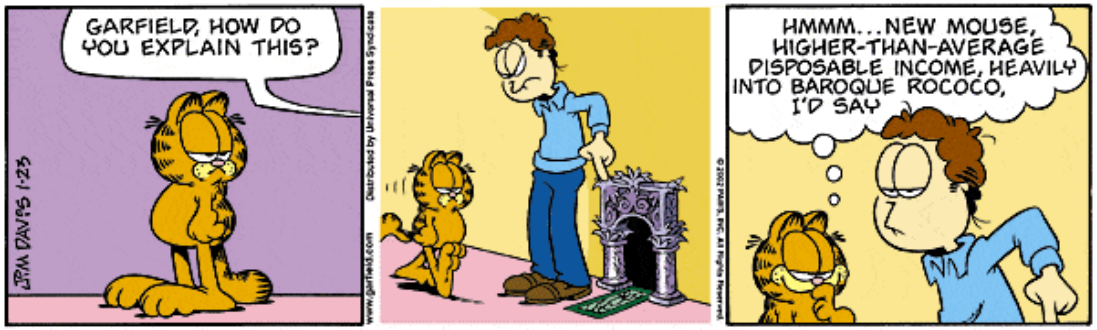
- Logaritmování:

$$\ln(Y_i) = \alpha + \beta_1 \ln(X_{1i}) + \dots + \beta_k \ln(X_{ki}),$$

kde $\alpha = \ln(\alpha_1)$.

- Přidáním náhodné složky $\rightarrow$ LRM (s logaritmy původních proměnných).

Interpretace parametrů



- Původní model (level-level): *„jestliže se X_j zvýší o jednotku, potom Y má tendenci zvýšit se o β_j jednotek (za předpokladu, že se hodnoty ostatních vysvětlujících proměnných se nemění)“*.
- Interpretace v jednotkách proměnných (dolary, tuny, apod.).
- Logaritmování = bezrozměrné veličiny.
- Log-log model: *elasticity*, tedy *„jestliže se X_j zvýší o jedno procento, potom má Y tendenci zvýšit se o β_j procent (za předpokladu, že se hodnoty ostatních vysvětlujících proměnných se nemění)“*.

Otázka logaritmování

- Pozor na logaritmy nul a záporných čísel!
- Log-level nebo Level-log model.
- Část proměnných v logaritmech a část ne.
- Příklad z ekonomie práce: závisle proměnná (Y) je logaritmus mzdy každého jednotlivce; vysvětlující proměnné počet let vzdělání (X_1) a počet let pracovních zkušeností (X_2).

$$\ln(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

- *„Jestliže se X_1 zvýší o jednotku, zvýší se závisle proměnná o $\beta_1 \cdot 100$ procent (za předpokladu, že se hodnoty ostatních vysvětlujících proměnných nemění).“*

Příklad mezd – rozšíření

- Pracovní zkušenosti nemají lineární vliv.
- Nová vysvětlující proměnná: druhá mocnina zkušeností.

$$\ln(Y_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{2i}^2 + \epsilon_i.$$

- Stále LRM!

Interakce proměnných

- Vztah mezi vysvětlujícími proměnnými.

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

- Třetí vysvětlující proměnnou $X_1 X_2$:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \epsilon_i.$$

- Stále LRM, nekonstantní vliv X_1 (a X_2) na Y.

$$Y_i = \alpha + [\beta_1 + \beta_3 X_{2i}] X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

- Mezní vliv X_1 na Y: $[\beta_1 + \beta_3 X_{2i}]$.
- Mezní vliv X_2 na Y: $[\beta_2 + \beta_3 X_{1i}]$.
- Obvykle mezní vliv vyhodnocován a prezentován v průměru pozorovaných dat (např. $\left[\beta_1 + \beta_3 \overline{X}_{2i}\right]$.)

Interakce proměnných – příklad

- Příklad vlivu vzdělání na mzdu.
  - Y = logaritmus mzdy;
  - X_1 = počet let vzdělání;
  - X_2 = skóre při testu inteligence.
- Mezní vliv X_1 na Y roven β_1: parametr je označován jako „výnosy ze vzdělání (the return to schooling)“.
- Nová proměnná odpovídající součinu vysvětlujících proměnných $\rightarrow$ $[\beta_1 + \beta_3 X_{2i}]$.
- Analýza jestli se výnosy ze vzdělání liší pro různé skupiny lidí.
- Mají inteligentní studenti větší užitek ze vzdělání než studenti méně inteligentní? (sledujeme statistickou významnost parametru β_3).

Rozhodování o nelineární podobě

- Ekonomická teorie $\times$ možno více specifikací.
- Příklad:

$$Y_i = \alpha + \beta_1 X_{1i} + \epsilon_i,$$
$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i.$$

- Využijeme t-test pro testování $H_0 : \beta_2 = 0$ nebo alternativně porovnání korigovaných koeficientů determinace, $\overline{R}^2$.
- Obecně: zkoušet různé specifikace (problém s multikolinearitou a zkreslením při opomenutí důležité vysvětlující proměnné).

Rozhodování o nelineární podobě

- Korigovaný koeficient determinace: modely se stejnou vysvětlovanou proměnnou!
- Jak rozhodnout mezi modely s různě transformovanými vysvětlovanými proměnnými? $\rightarrow$ nelehká otázka.
- Speciální případ:

$$Y_i = \alpha + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \epsilon_i,$$
$$\ln(Y_i) = \alpha + \beta_1 \ln(X_{1i}) + \ldots + \beta_k \ln(X_{ki}) + \epsilon_i.$$

- První model – lineární regrese; druhý model – log-lineární regrese (nezáleží jestli jsou všechny nebo jen část vysvětlujících proměnných vyjádřena v podobě logaritmů).
- Obě vysvětlované proměnné nejsou přímo srovnatelné.

Rozhodování o logaritmu vysvětlované proměnné

- Nová proměnná: $Y_i^* = \frac{Y_i}{\tilde{Y}}$, kde $\tilde{Y}$ je geometrický průměr nelogaritmované závisle proměnné (srovnatelné).
- SSR_{LIN} a SSR_{LOG} pro lineární a log-lineární regresi s použitím závisle proměnných Y^* a $\ln(Y^*)$.
- Pokud $SSR_{LIN} > SSR_{LOG} \rightarrow$ testová statistika hypotézy, že lineární a log-lineární regrese vyrovnávají data stejně:

$$LL_1 = N \ln\left(\frac{SSR_{LIN}}{SSR_{LOG}}\right).$$

- Pokud $SSR_{LOG} > SSR_{LIN}$:

$$LL_2 = N \ln\left(\frac{SSR_{LOG}}{SSR_{LIN}}\right).$$

- Rozdělení χ_1^2 (kritická hodnota jednostranného testu na hladině významnosti 5 % je 3.841) $\rightarrow$ v případě zamítnutí $\rightarrow$ preference modelu s nižším SSR.

Sargan (1964)

- M_1 model lineární a M_2 log-lineární.
- Za předpokladu nezávislých a normálně rozdělených náhodných složek $\rightarrow$ OLS odhady rozptylů reziduí obou modelů $\widehat{\sigma}^2_{M_1}$ a $\widehat{\sigma}^2_{M_2}$.
- Sarganovo kritérium:

$$S = \left(\frac{\widehat{\sigma}_{M_1}}{g\widehat{\sigma}_{M_2}} \right)^N,$$

  kde N počet pozorování a g je geometrický průměr vysvětlovaných proměnných $y_1, \ldots, y_N$.
- Pokud $S < 1$, potom data hovoří ve prospěch M_1.
- Pokud $S > 1$, potom data hovoří ve prospěch M_2.

BM test

- Bera a McAleer (1989).

$$\log y_t = \beta_0 + \beta_1 x_t + u_{0t},$$
$$y_t = \beta_0 + \beta_1 x_t + u_{1t}.$$

1. Získání vyrovnaných hodnot $\log \widehat{y}_t$ a $\tilde{y}_t$. Vyrovnaná hodnota y_t z log-lineární rovnice je $\exp(\log \widehat{y}_t)$. Predikovaná hodnota $\log y_t$ z lineární rovnice je $\log \tilde{y}_t$.
2. Odhad pomocných regresí a získání reziduí $\widehat{v}_{1t}$ a $\widehat{v}_{0t}$:

$$\exp(\log \widehat{y}_t) = \beta_0 + \beta_1 x_t + v_{1t}, \quad \log \tilde{y}_t = \beta_0 + \beta_1 x_t + v_{0t}.$$

3. Standardní t-test parametrů θ_0 a θ_1 v pomocných regresích:

$$\log y_t = \beta_0 + \beta_1 x_t + \theta_0 \widehat{v}_{1t} + w_{0t}, \quad y_t = \beta_0 + \beta_1 x_t + \theta_1 \widehat{v}_{0t} + w_{1t}.$$

4. Pokud $\theta_0 = 0$ není zamítnuta, volíme log-lineární podobu. Pokud $\theta_1 = 0$ nezamítnuta, volíme lineární model (problém, pokud současně zamítáme nebo nezamítáme obě hypotézy).

PE test

- MacKinnon, White a Davidson (1983).
- První krok stejný jako u BM testu.
- Ve druhém kroku analogický test $\theta_0 = 0$ a $\theta_1 = 0$ v umělých regresích:

$$\log y_t = \beta_0 + \beta_1 x_t + \theta_0[\tilde{y}_t - \exp(\log \widehat{y}_t)] + \epsilon_{0t},$$
$$y_t = \beta_0 + \beta_1 x_t + \theta_1[\log \widehat{y}_t - \log \tilde{y}_t] + \epsilon_{1t}.$$

- Existence dalších testů (zejména využívajících Box-Coxovu transformaci a metodu maximální věrohodnosti).

Obsah tématu

Změna měřítka závisle proměnné

- Násobíme vysvětlovanou proměnnou konstantou c (nenulovou):
  1. OLS odhady úrovňová konstanta a parametry sklonu násobeny konstantou c.
  2. Nemění se koeficient determinace, R^2.
  3. Směrodatné odchylky odhadů všech parametrů násobeny c.
  4. Součet čtverců reziduí násoben c^2 (rezidua se zvyšují c krát).
  5. Směrodatná odchylka reziduí násobena c krát.
  6. Nemění se výsledky testů statistické významnosti parametrů (t-testy, F-test).

Změna měřítka nezávisle proměnné

- Násobení vysvětlující proměnné konstantou c (nenulovou):
  1. OLS odhad parametru sklonu dělen konstantou c (násoben c^{-1}).
  2. Nemění se koeficient determinace, R^2.
  3. Směrodatná odchylka odhadu jen měněného parametru dělena c (násobena c^{-1}).
  4. Součet čtverců reziduí nezměněn.
  5. Směrodatná odchylka reziduí nezměněna.
  6. Nemění se výsledky testů statistické významnosti parametrů (t-testy, F-test).

RESET test

- Testování chybné specifikace modelu (Regression Specification Error Test).
- Detekce opomenutých proměnných a nekorektní funkční podoby.

RESET test – modelové vyjádření

- Princip postupu:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i,$$
$$\widehat{Y}_i = \widehat{\beta}_1 + \widehat{\beta}_2 X_{i2} + \widehat{\beta}_3 X_{i3}.$$

- Předpoklad dvou umělých modelů

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \gamma_1 \widehat{Y}_i^2 + \epsilon_i,$$
$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \gamma_1 \widehat{Y}_i^2 + \gamma_2 \widehat{Y}_i^3 + \epsilon_i.$$

RESET test – princip

- Test chybné specifikace v prvním případě: $H_0 : \gamma_1 = 0$, $H_1 : \gamma_1 \neq 0$.
- Test chybné specifikace ve druhém případě: $H_0 : \gamma_1 = \gamma_2 = 0$, $H_1 : \gamma_1 \neq 0$ a (nebo) $\gamma_2 \neq 0$.
- První případ t-test nebo F-test; druhý případ F-test.
- Zamítnutí H_0 = model neadekvátní a měl by být zlepšen; nezamítnutí H_0 = test nebyl schope detekovat chybnou specifikaci.
- Princip: $\widehat{Y}_i^2$ a $\widehat{Y}_i^3$ jsou polynomiální funkce X_{i2} a X_{i3} $\rightarrow$ druhá a třetí mocnina rovnice vyrovnaných hodnot obsahuje mocniny a křížové členy vysvětlujících proměnných $\rightarrow$ polynomy aproximují různé funkční formy $\Rightarrow$ nekorektní původní model = zahrnutí $\widehat{Y}_i^2$ a $\widehat{Y}_i^3$ zvýší kvalitu vyrovnání.
- Podobně problém nezahrnutí proměnných: pokud korelovány s X_{i1} a X_{i2} = korelovány pravděpodobně i s jejich mocninami (vyřešení zahrnutím $\widehat{Y}_i^2$ a $\widehat{Y}_i^3$).

RESET test – shrnutí

- Pokud významně zkvalitníme model zahrnutím umělým zahrnutím predikovaných hodnot modelem, musí být původní model neadekvátně specifikován.
- Test přímo neříká, co dělat dál.
- Užitečný pro zjištění slabě specifikovaných modelů.
- Ne vždy rozhodne mezi alternativními modely (RESET nemusí zamítnout žádnou z alternativ).
- Zobecnění přidáním vyšších mocnin vyrovnaných hodnot (testování výraznějších nelinearit).

Obsah tématu

Obsah tématu

Motivace

- Komplikovanější než F-test $\times$ širší uplatnění.

- Věrohodnostní funkce: $L\left(\alpha, \beta_1, \ldots, \beta_k, \sigma^2\right)$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - \alpha - \beta_1 X_{1i} - \ldots - \beta_k X_{ki}\right)^2\right]$$

$$= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{N}\left(Y_i - \alpha - \beta_1 X_{1i} - \ldots - \beta_k X_{ki}\right)^2\right].$$

Východiska

- ML odhady parametrů odpovídají OLS odhadům: $\widehat{\alpha}$, $\widehat{\beta}_1$,. . . , $\widehat{\beta}_k$.
- ML odhad rozptylu náhodných složek není nestranný:

$$\widehat{\sigma}^2 = \frac{\sum \left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} \ldots \widehat{\beta}_k X_{ki} \right)^2}{N}$$
$$= \frac{\sum \widehat{\epsilon}_i^2}{N}.$$

- Hodnota věrohodnostní funkce pro neomezený model (MLE):

$$L\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U, \widehat{\sigma}^{2U}\right).$$

- Věrohodnostní funkce vyhodnocená v odhadech omezeného modelu (MLE):

$$L\left(\widehat{\alpha}^R, \widehat{\beta}_1^R, \ldots, \widehat{\beta}_k^R, \widehat{\sigma}^{2R}\right).$$

Ilustrace

- Regresní model se třemi vysvětlujícími proměnnými a hypotéza

$$H_0 : \beta_1 = 0, \quad \beta_2 + \beta_3 = 1.$$

- Zohledněním omezení z nulové hypotézy získáme omezený model

$$Y_i - X_{2i} = \alpha + \beta_3 (X_{3i} - X_{2i}) + \epsilon_i.$$

- OLS odhady $\rightarrow \widehat{\alpha}^R$ a $\widehat{\beta}_3^R$.
- Hodnoty $\widehat{\beta}_1^R$ a $\widehat{\beta}_2^R$? $\rightarrow$ omezení plynoucí z H_0, tedy $\widehat{\beta}_1^R$ a $\widehat{\beta}_2^R = 1 - \widehat{\beta}_3^R$.
- Testy věrohodnostního poměru i pro hypotézy zahrnující nelineární restrikce: např. $H_0 : \beta_1 = \beta_2^3,$, $\beta_3 = \frac{1}{\beta_2} \rightarrow$ obecně $H_0 : g(\beta_1, \ldots, \beta_k) = 0$, kde $g(\cdot)$ je množina až k nelineárních funkcí.
- Odhad nelineárních modelů v ekonometrických programech.

Test věrohodnostního poměru

- Věrohodnostní poměr:

$$\lambda = \frac{L\left(\widehat{\alpha}^R, \widehat{\beta}_1^R, \ldots, \widehat{\beta}_k^R, \widehat{\sigma}^{2R}\right)}{L\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U, \widehat{\sigma}^{2U}\right)}.$$

- Testová statistika je $-2\ln(\lambda)$.
- Rozdělení této statistiky (aproximativně): $-2\ln(\lambda) \sim \chi^2_q$ (q je počet omezení obsažených v H_0).
- Intuice: zavedení restrikcí vede k nižší hodnotě věrohodnostní funkce.
- Platí: $L\left(\widehat{\alpha}^R, \widehat{\beta}_1^R, \ldots, \widehat{\beta}_k^R, \widehat{\sigma}^{2R}\right) \leq L\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U, \widehat{\sigma}^{2U}\right)$ a tedy $0 \leq \lambda \leq 1$.
- H_0 pravdivá $\Rightarrow$ λ bude velmi blízko 1 $\Rightarrow$ testová statistika $-2\ln(\lambda)$ malá.

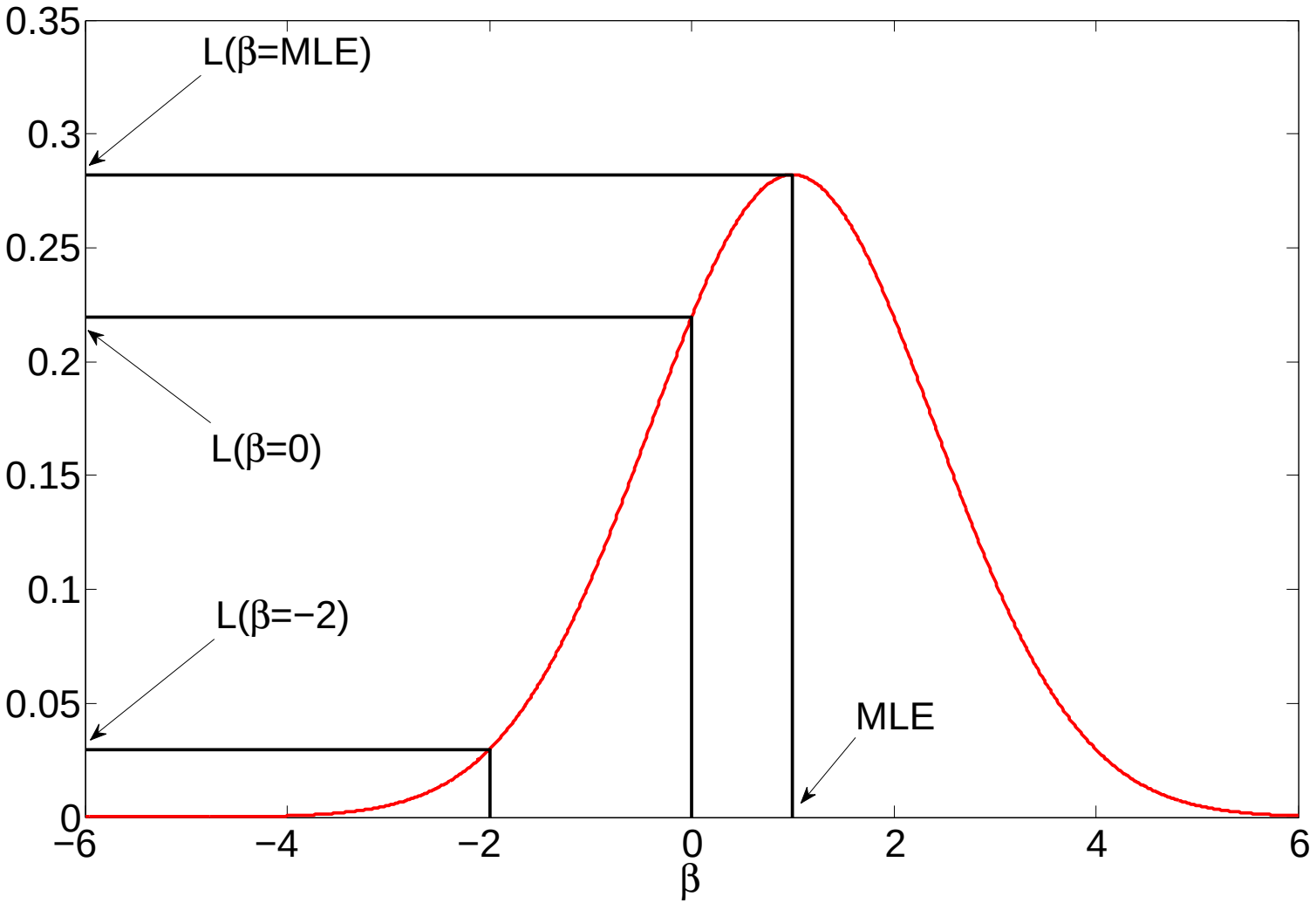
Příklad 1

- Jednoduchý regresní model se známým rozptylem, bez úrovňové konstanty a jediným koeficientem, β.
- Neomezená věrohodnostní funkce: $L(\beta)$.
- Test hypotézy $H_0 : \beta = 0 \Rightarrow$ omezená věrohodnostní funkce $L(\beta = 0)$.
- Obrázek: $N(2, 1)$.
- Věrohodnostní poměr:

$$\lambda = \frac{L(\beta = 0)}{L(\beta = MLE)}.$$

- $\lambda = 0.773$ a $-2\ln(\lambda) = 0.515 \rightarrow$ kritická hodnota pro χ^2_1 je 3.84 (jednostranný test a tudíž hodnota 3.84 odpovídá 95% kvantilu daného rozdělení) $\rightarrow$ nezamítáme nulovou hypotézu, že $\beta = 0$.

Věrohodnostní funkce



Příklad 2

- Hypotéza $H_0 : \beta = -2 \rightarrow L(\beta = -2)$
- Hodnota věrohodnostní funkce je zde mnohem nižší než hodnota v MLE.
- Věrohodnostní poměr:

$$\lambda = \frac{L(\beta = -2)}{L(\beta = MLE)} = \frac{0.031}{0.282} = 0.110.$$

- $-2\ln(\lambda) = 4.416 \rightarrow$ 5% kritická hodnota 3.84 $\rightarrow$ zamítáme nulovou hypotézu, že $\beta = -2$.

Alternativa pro LRM

- Věrohodnostní funkce pro model vícenásobné regrese:

$$L\left(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k, \widehat{\sigma}^2\right)$$
$$= \frac{1}{\left(2\pi\widehat{\sigma}^2\right)^{\frac{N}{2}}} \exp\left[-\frac{1}{2\widehat{\sigma}^2}\sum_{i=1}^{N}\left(Y_i - \widehat{\alpha} - \widehat{\beta}_1 X_{1i} - \ldots - \widehat{\beta}_k X_{ki}\right)^2\right].$$

- Po dosazení výrazu pro odhad rozptylu:

$$L\left(\widehat{\alpha}, \widehat{\beta}_1, \ldots, \widehat{\beta}_k, \widehat{\sigma}^2\right) \propto \frac{1}{\left(\widehat{\sigma}^2\right)^{\frac{N}{2}}} \propto \frac{1}{\left(SSR\right)^{\frac{N}{2}}},$$

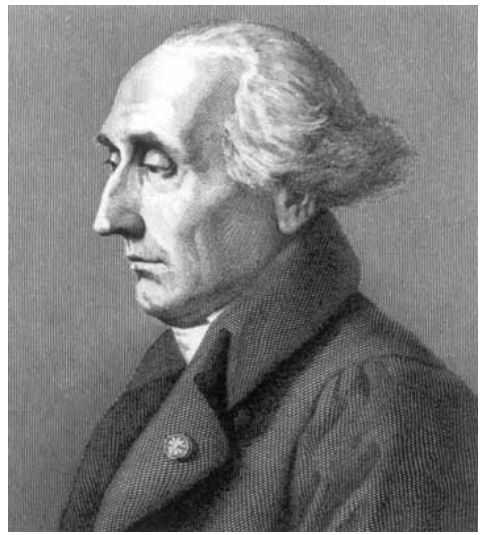
kde $SSR = \sum \widehat{\epsilon}_i^2$.

- Věrohodnostní poměr: $\lambda = \dfrac{\dfrac{1}{\left(SSR^R\right)^{\frac{N}{2}}}}{\dfrac{1}{\left(SSR^U\right)^{\frac{N}{2}}}} = \left(\dfrac{SSR^U}{SSR^R}\right)^{\frac{N}{2}}.$

Waldův test a test Lagrangeových multiplikátorů

- Varianty testů založených na věrohodnostním poměru.

Abraham Wald
(1902–1950)



Joseph-Louis Lagrange
(1736–1813)

Waldův test

- Odhad pouze neomezeného modelu.
- Příklad: hypotéza $H_0 : g(\alpha, \beta_1, \beta_2, \ldots, \beta_k) = c$
- ML odhady $\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U$.
- Idea: v případě správnosti hypotézy H_0 odhady v blízkosti hodnot splňujících omezení.
- Mělo by platit: $g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U)$ nebude příliš vzdálené od hodnoty c.

Waldův test – dokončení

- Waldova statistika:

$$W = \frac{\left[g\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U\right) - c\right]^2}{var\left[g\left(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U\right)\right]}.$$

- Jmenovatel někdy snadno spočítatelný, např. pro $g(\widehat{\alpha}^U, \widehat{\beta}_1^U, \ldots, \widehat{\beta}_k^U) = \widehat{\beta}_1^U + \widehat{\beta}_2^U$:

$$var\left(\widehat{\beta}_1^U + \widehat{\beta}_2^U\right) = var\left(\widehat{\beta}_1^U\right) + var\left(\widehat{\beta}_2^U\right) + 2cov\left(\widehat{\beta}_1^U, \widehat{\beta}_2^U\right).$$

- Pro případ nelineárních restrikcí nutné komplikovanější statistické metody $\rightarrow$ dokáží spočítat ekonometrické balíčky.
- Rozdělení testové statistiky:

$$W \sim \chi_q^2,$$

kde q je počet omezení v rámci nulové hypotézy.

Test Lagrangeových multiplikátorů

- Odhad pouze omezeného modelu.
- Příklad: neomezený model jednoduchý regresní model s jediným koeficientem, β; omezený model v rámci hypotézy $H_0 : \beta = c$.
- $\widehat{\beta}^R = c$.
- Motivace testu: v případě platnosti H_0 maximálně věrohodný odhad omezeného modelu by neměl být příliš vzdálen od ML odhadu neomezeného modelu (pro náš příklad by c nemělo být příliš vzdálené od $\widehat{\beta}$, tedy OLS (ML) odhadu).
- Diferenciální počet říká, že v maximu věrohodnostní funkce je první derivace funkce nulová (což odpovídá směrnici tečny v bodě).
- Pokud je H_0 pravdivá, měla by být derivace věrohodnostní funkce vyhodnocená v $\widehat{\beta}^R$ blízko nule.

Test Lagrangeových multiplikátorů – dokončení

- Testová statistika:

$$LM = \frac{\left[\mathrm{d} \ln L\left(\widehat{\beta}^R\right)\right]^2}{I\left(\widehat{\beta}^R\right)}.$$

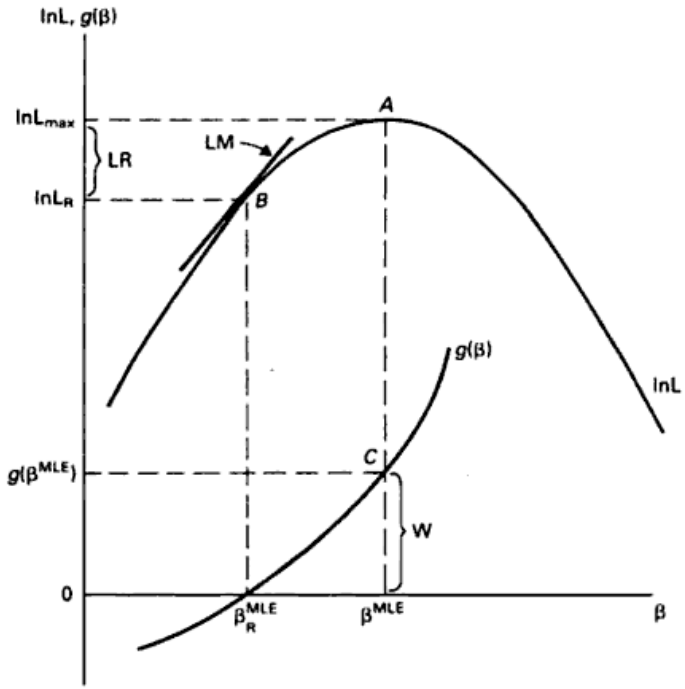
- Intuitivně: jak hodně vzdálený nule je sklon tečny věrohodnostní funkce při zohlednění restrikcí.
- Čitatel počítá směrnici tečny v tomto bodě $\times$ velikost odchylky vyjádřena relativně vzhledem k nejistotě spojenou s tímto odhadem.
- Jmenovatel LM je vstažen k nejistotě odhadu: $I(\cdot)$ je obecně tzv. *informační matice* (vyhodnocená v omezeném odhadu, její inverze odpovídá kovarianční matici).
- LM statistika má rozdělení, aproximativně (asymptoticky) chí-kvadrát:

$$LM \sim \chi^2_q,$$

kde q je počet restrikcí v kontextu H_0.

Porovnání testů – obrázek

- Likelihood ratio (LR) test, Waldův test (W), test Lagrangeových multiplikátorů (LM).
- Log-likelihood ($\ln L$) jako funkce β; β^{MLE} maximum; omezení $g(\beta) = 0$; hodnota β_R^{MLE}.



Zdroj: Kennedy (2008) – A Guide to Econometrics.

Porovnání testů

- LR test: omezení pravdivé $\Rightarrow \ln L_R$ (maximum $\ln L$ při omezeních) by nemělo být statisticky menší než $\ln L_{\max}$ (neomezené maximum). Test nulovosti vertikální vzdálenosti $(\ln L_{\max} - \ln L_R)$.
- W test: omezení $g(\beta) = 0$ pravdivé $\Rightarrow g(\beta^{MLE})$ by nemělo být statisticky menší než 0. Test nulovosti vertikální vzdálenosti $g(\beta^{MLE})$ od nuly (naše omezení) resp. nulovosti horizontální odchylky β^{MLE} od β_R^{MLE}.
- LM test: sklon $\ln L$ v maximu (vzhledem k β) nulový $\rightarrow$ omezení pravdivé $\Rightarrow$ sklon $\ln L$ v omezeném odhadu β_R^{MLE} nevýznamně vzdálený od nuly.
- *Statistiky pro test věrohodnostního poměru, Waldův test a test Lagrangeových multiplikátorů jsou asymptoticky ekvivalentní.*

Konec