Nastínění problému

Lineární regrese se mnoho používá, ale málo se již ví, že má některé předpoklady, které by pro její nasazení měly být splněny.

Jedním z těchto předpokladů je normalita chyb. Každé měření je zatíženo nějakou chybou. Pro pro model obyčejné (Ordinary least squares) regrese to není problém do doby, dokud jsou chyby v měřeních opravdu náhodné.

Korelované chyby aka. Autoregrese

Že máme v datech korelované chyby neboli autoregresi znamená, že mezi jednotlivými chybami lze vystopovat nějakou korelaci.

V praxi často uvažujeme pouze to, že i-tá chyba byla ovlivněna měřením $i-1$. Jinak řečeno aktuální chyba byla ovlivněna předchozí a má vliv pouze na následující.

V takovýchto datech nás potom zajímá *autoregresní parametr*. Když už zhřešíme a použijeme OLS na data trpící autoregresí, můžeme takového modelu alespoň získat z autoregresní parametr. Stačí jen vypočítat korelaci mezi *rezidui* [1] s indexy $1, ..., n-1$ a $2, ..., n$.

Takováto korelace mezi chybami bývá pěkně vidět na grafech *teoretické vs. navzorkované kvantily*, *indexy vs. rezidua* a $rezidua\ r_1, ..., r_{n-1}\ vs.\ reziduar_2, ..., r_n$. Každý graf bude vysvětlen v příslušné pasáži.

Jak to spravit?

Zobecněná metoda nejmenších čtverců

Jednou z možností je použití metody Generalised leased squares, která v jistých ohledech přirozeně rozšiřuje metodu Ordinary least squares. Ve zkratce ve výpočtu regresního modelu figuruje nějaká *kovariační matice náhodných chyb*, ve které jsou reprezentovány rozptyly v datech (rozptyly jsou alfou a omegou těchto metod). U standardního OLS na tuto matici klademe požadavek, aby byla *diagonální*. Pokud je ale v datech autoregrese, diagonalita matice je často porušena.

Řešit to lze tak, že do *kovariační matice* začleníme vhodným způsobem *autoregresní parametr*, který jsme získali například z nešťastného použití

[1] Reziduum je vzdálenost jednotlivého bodu od regresní přímky, která body protíná

OLS. Výsledkem je potom model, který tuto autoregresi, tedy korelované chyby v měření, bere v potaz.

Ve vizualizaci je přímka tohoto modelu zobrazená zeleně.

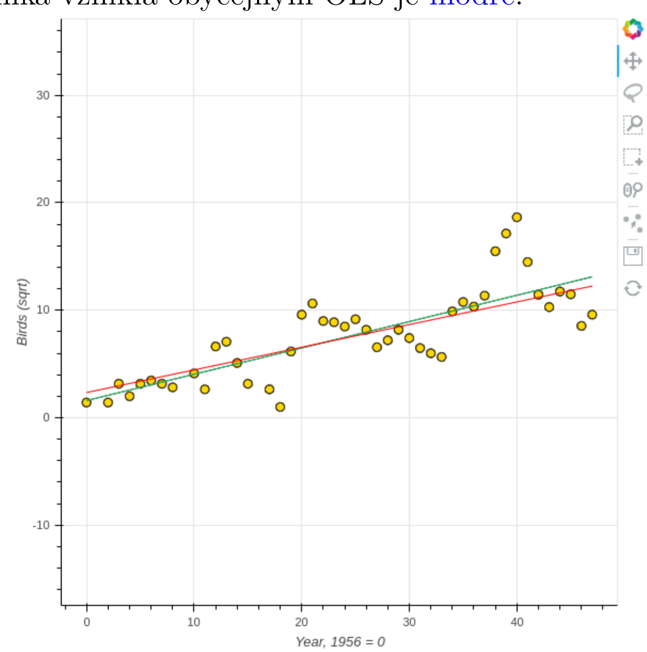
Cochrannova-Orcuttova metoda

Další možnost se nazývá Cochrannova-Orcuttova metoda. Její princip je v celku jednoduchý. Odhadneme autoregresní parametr, upravíme data (odečteme tu část chyby, která má být korelovaná s předchozí) a následně použijeme *znovu OLS* na těchto upravených datech. Tento postup opakujeme, dokud nedokonvergujeme k výsledku. V praxi často konvergence končí už po první iteraci a proto jsem si dovolil i ve svojí vizualizaci po první iteraci skončit.

Přímka vzniklá z Chorannovy-Orcuttovy metody je zobrazena červeně.

Vizualizace

Originální dataset je zobrazen žlutými puntíčky. Pro úplnost dodám, že přímka vzniklá obyčejným OLS je modře.



K manipulaci s datasetem slouží *toolbox* v pravém horním rohu.

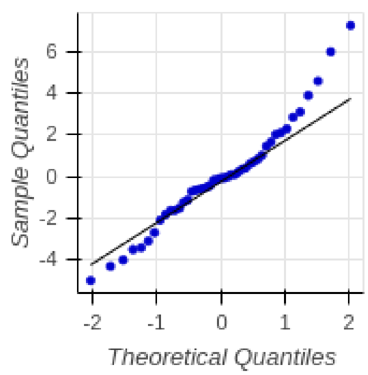
Point draw tool (ikonka tří teček s ukazatelem) slouží k přidávání bodů (prosté kliknutí), označování bodů, jejich přesunu (tažením) a odstranění (klávesou *backspace*). Více bodů lze označit držením klávesy *shift*. Ekvivalentní a možná pohodlnější možností je zvolit body *lasem* nebo *čtvercovým výběrem*. Po označení bodů je však nutno pro manipulaci s nimi zase kliknout na *Point draw tool*!

Model se přepočítává hned po změně datasetu.

Theoretické kvantily vs. navzorkované kvantily

Rezidua by měla mít normální rozdělení. Tento graf zobrazuje, jak velká by měla být dokonale normální rezidua (čára) oproti tomu, jak velká jsou rezidua ve skutečnosti.

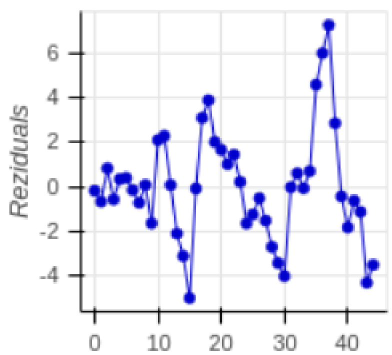
Lze vidět, že na pravé straně grafu se vychylují výš a na levá straně níž. To znamená, že je tam nepoměr mezi aktuálními rezidui a tím, jak by rezidua vypadala, kdyby opravdu pocházela z normálního rozdělení.

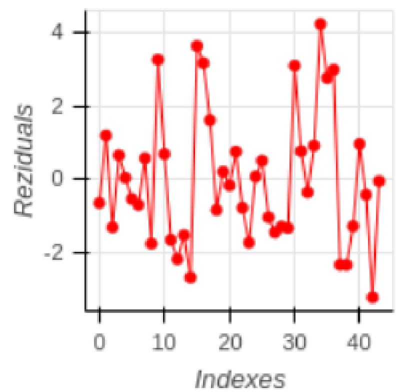


Indexy vs. rezidua

Tento graf jednoduše zobrazuje v řadě za sebou jednotlivá rezidua. Je patrné, že na začátku je v datech malý rozptyl, tudíž rezidua jsou malá a tento rozptyl se postupně zvyšuje a na konci jsou rezidua až o velikosti 6.

Ideálně by rezidua v tomto grafu měla být rozprostřena na ose y náhodně. Lepší splnění této podmínky je vidět na grafu reziduí Cochrannovy-Orcuttovy





metody.

Rezidua $r_1, ..., r_{n-1}$ vs. rezidua $r_2, ..., r_n$

Jakkoliv složitě tento název zní, jeho interpretace je nejintuitivnější a opravdu tak prostá. Jedná se o zobrazení o jedničku posunutých reziduí na sebe, tudíž do grafu se vykreslí body [první reziduum; druhé reziduum], [druhé reziduum; třetí reziduum], ...

Zde se nejkrásněji ukáže korelace mezi rezidui, protože pokud existuje, budou takto zobrazené body vykazovat stoupavou nebo klesavou tendenci. Pokud jimi pak protneme přímku, bude vychýlená od osy x.

V ideálním případě by měly být body na tomto grafu náhodně rozesety a přímka rovnoběžná s osou x.

Opět pro porovnání přidávám obrázek z vizualizace.