9 Paskaita. Paprasto tiesinio regresinio modelio tyrimas.

9.1 Modelio netinkamumo priežastys.

Tarkime, nagrinėjame intervalinių kintamųjų X ir Y tiesinį regresinį modelį. Išvardinsime svarbiausias priežastis, dėl kurių regresinis modelis galėtų netikti:

1. Duomenyse yra išskirčių.
2. Kintamųjų priklausomybė nėra tiesinė.
3. Stebėjimai yra hereroskedastiniai (esant skirtingoms X reikšmėms Y dispersijos skiriasi).
4. Paklaidų skirstiniai nėra normalieji.
5. Paklaidos priklausomos.

Negalima suabsoliutinti nė vienos iš šių galimų priežasčių.

9.2 Išskirtys.

Išskirtys yra labai stipriai nuo kitų duomenų besiskiriantys duomenys. Jie gali labai stipriai pakeisti regresijos tiesės parametrus. Norėdami įsitikinti, kad taip yra, galime apskaičiuoti regresijos parametrus su išskirtimi ir be jos. Išskirtys iš duomenų šalinamos tik tuomet, jei tam yra pakankamai pagrindo, pvz., duomenyse yra klaida, arba, jei duomenys buvo paveikti vienkartinių nebūdingų veiksnių. Taikomi keli alternatyvūs išskirčių nustatymo metodai.

1. Stebėjimo įtakos indeksas vertina, tik tai, kiek toli nuo $\bar{x}$ yra stebėjimas x_j. Stebėjimo (x_j, y_j) įtakos indeksas skaičiuojamas taip:

$$h_j = \frac{1}{n} + \frac{(x_j - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \bar{x} = \sum_{i=1}^{n} x_i/n.$$

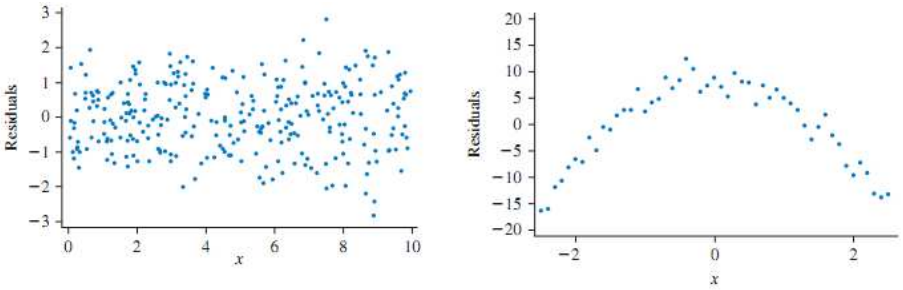
   Galima pateikti tokią "grubią" rekomendaciją: stebėjimas (x_j, y_j) gali būti laikomas išskirtimi, jei $h_j > 4/n$.

2. Standartizuotos liekanos. Tai yra liekamųjų paklaidų ε_i z-reikšmės. T.y. reikia iš ε_i atimti jos vidurkį ir padalinti iš standartinio nuokrypio. Kadangi visų ε_i vidurkiai lygūs nuliui, standartizuota liekana lygi

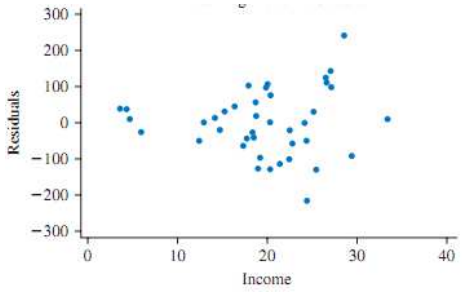
$$SR_i = \frac{\hat{\varepsilon_i}}{\sqrt{MSE(1-h_i)}}, MSE = SSE/(n-2), SSE = \sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

   Stebėjimą (x_j, y_j) laikome išskirtimi, jei $|SR_i| > 3$.

3. Kuko matas.



1 pav.: Dviejų modelių liekamosios paklaidos



2 pav.: Heteroskedastinės paklaidos

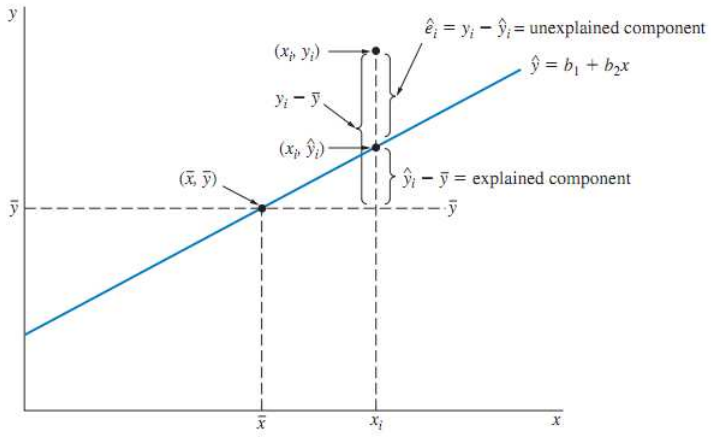
9.3 Liekamųjų paklaidų grafikai.

Grafinei prielaidų diagnostikai paprastai naudojami liekamųjų paklaidų arba standartizuotų liekanų grafikai. Jeigu regresijos modelis tinka, visi $\hat{\varepsilon}_i$ turėtų būti labai nedideli ir daugmaž vienodai išsibarstę apie tiesę $y = 0$. Pav. 1 kairėje yra tinkamo regresinio modelio paklaidos, o dešinėje – blogai parinkto modelio paklaidos, nes paklaidos nėra atsitiktinai išsibarsčiusios. Iš sklaidos diagramų galima taip pat spręsti apie duomenų heteroskedastiškumą. Heteroskedastinės liekanos pavaizduotos pav. 2.

Stebėjimų normalumą galima įvertinti nubraižius liekamųjų paklaidų histogramą, kvantilių atitikimo grafiką, pritaikius Šapiro-Wilk arba kitą normalumo tikrinimo kriterijų.

9.4 Priklausomų paklaidų problema.

Kartais regresijos liekamosios paklaidos būna priklausomos. Dažnai šia savybe pasižymi daugelis ekonominių indeksų ir rodiklių, pvz., akcijų kursai, BVP vienam gyventojui, nedarbingumo lygis, užimtumo lygis ir pan. Regresiniai modeliai su koreliuojančiomis liekanomis nagrinėjami laikinių sekų teorijoje. Norėdami nustatyti, ar paklaidos koreliuoja, naudojame Durbino-Watsono kriterijų. Jis leidžia nustatyti, ar yra vadinamasis *autoregresijos* modelis.



3 pav.: Paaiškinamoji ir nepaaiškinamoji y_i komponentės

Prielaida: regresijos modelio paklaidas sieja ryšys

$$\varepsilon_i = \rho\varepsilon_{i-1} + z_i,$$

čia $z_i \sim \mathcal{N}(0, \sigma^2)$, ir $z_1, z_2, \ldots$ yra nepriklausomi atsitiktiniai dydžiai. Paklaidos ε_i nekoreliuoja, kai $\rho = 0$. Patikrinsime šią hipotezę.

$$\begin{cases} H_0 : \rho = 0, \\ H_1 : \rho \neq 0. \end{cases}$$

Kriterijaus statistika $d = \dfrac{\sum_{i=2}^{n} (\hat{\varepsilon}_i - \hat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \hat{\varepsilon}_i^2}$ kinta nuo 0 iki 4. Kuo d arčiau 2, tuo mažiau tikėtina, kad autokoreliacija yra.

9.5 Determinacijos koeficientas.

Regresijos tiesė visada eina per vidurkių tašką $(\bar{x}, \bar{y})$. Kiekvieną y_i ir $\bar{y}$ skirtumą galima išskaidyti į dvi dalis (žr. pav. 3).

$$y_i - \bar{y} = (y_i - \hat{y}(x_i)) + (\hat{y}(x_i) - \bar{y}) = \varepsilon_i + (\hat{y}(x_i) - \bar{y}), \tag{1}$$

čia $\hat{y}(x_i)$ yra x_i atitinkanti y reikšmė, apskaičiuota iš regresijos lygties. Pakėlę abi lygties (1) puses kvadratu ir sudėję, gausime tokią lygybę:

$$SST = SSR + SSE,$$

čia $SST = \sum_{i=1}^{n} (y_i - \bar{y})^2, SSR = \sum_{i=1}^{n} (\hat{y}(x_i) - \bar{y})^2, SSE = \sum_{i=1}^{n} \varepsilon_i^2$.

Suma SST vadinama *visa kvadratų suma*, SSR – *regresijos kvadratų suma*, SSE – *liekamųjų paklaidų kvadratų suma*. SST vertina, kaip y_i reikšmės yra išsibarsčiusios apie tiesę $y = \bar{y}$. SSE vertina, kaip y_i reikšmės yra išsibarsčiusios apie

regresijos tiesę $\hat{y}(x) = b_1 + b_2 x$. SSR – tai kvadratų suma, rodanti, kiek regresijos tiesė skiriasi nuo tiesės $y = \bar{y}$. Santykis SSR/SST vadinamas *determinacijos koeficientu* ir žymimas r^2. Determinacijos koeficientą galima interpretuoti, kaip santykį variacijos dalies, kurį paaiškina regresijos modelis ir visos Y variacijos. Kuo didesnė r^2 reikšmė, tuo geriau regresijos modelis aprašo duomenis. Jei visi duomenys yra regresijos tiesėje, $SSE = 0, SSR = SST, r^2 = 1$. Determinacijos koeficientas yra glaudžiai susijęs su Pirsono koeliacijos koeficientu ρ_{XY}. Paprastosios tiesinės regresijos atveju (kai yra vienas nepriklausomas kintamasis), determinacijos koeficientas r^2 sutampa su Pirsono koeliacijos koeficiento kvadratu. Didesnis r^2 reiškia, kad duomenys yra labiau koncentruoti apie regresijos tiesę. Tačiau, remiantis vien tik determinacijos koeficientu negalime pasakyti, ar modelis yra tinkamas. Kartais, ypač daugialypėje regresijoje, naudojamas *pataisytas determinacijos koeficientas* r^2_{adj}, kuris atsižvelgia į imties didumą ir nepriklausomų kintamųjų skaičių:

$$r^2_{adj} = 1 - (1 - r^2)\frac{n-1}{n-2}.$$

9.6 Regresijos parametrų pasikliautinieji intervalai

Gali atsitikti taip, kad paklaidos tenkina reikalavimus 1.–4., tačiau nepriklausomas kintamasis kintamajam Y įtakos neturi, bes jo koeficientas lygus nuliui, t.y. $y_i = a + 0 \cdot x_i + \varepsilon_i$. Dėl to būtina įvertinti regresijos koeficientų reikšmingumą.

Tarkime, kad paprastajam regresiniam modelyje $y_i = a + b \cdot x_i + \varepsilon_i$ galioja visos prielaidos liekamosioms paklaidoms. Modelyje yra du nežinomi parametrai a ir b ir σ^2. Anksčiau radome a ir b taškinius įverčius $\hat{a}$ ir $\hat{b}$. Galima apskaičiuoti parametrų a ir b pasikliautinuosius intervalus.

Remiantis modelio prielaidomis įrodyta, kad atsitiktinių dydžių a ir b skirstiniai yra normalieji. Be to,

$$E\hat{a} = a, D\hat{a} = \sigma^2\left(\frac{1}{n} + \frac{(\bar{x})^2}{(n-1)s_X^2}\right), E\hat{b} = b, D\hat{b} = \frac{\sigma^2}{(n-1)s_X^2}. \qquad (2)$$

Čia s_X^2 yra $x_1, \ldots, x_n$ empirinė dispersija. Pakeitę formulėse (2) σ^2 jos įverčiu $MSE = SSE/(n-2)$, gausime $\hat{a}$ ir $\hat{b}$ dispersijų įverčius:

$$s_a^2 = MSE\left(\frac{1}{n} + \frac{(\bar{x})^2}{(n-1)s_X^2}\right), s_b^2 = \frac{MSE}{(n-1)s_X^2}. \qquad (3)$$

Statistikos $\dfrac{\hat{a}-a}{s_a}, \dfrac{\hat{b}-b}{s_b}$ turi Stjudento skirstinius su $n-2$ laisvės laipsiais. Be to, SSE/σ^2 turi χ^2 skirstinį su $n-2$ laisvės laipsiais. Koeficientų a ir b ir dispersijos σ^2 pasikliautinieji intervalai skaičiuojami pagal formules:

$$\hat{a} \pm s_a t_{(1-Q)/2}(n-2), \quad \hat{b} \pm s_b t_{(1-Q)/2}(n-2), \left(\frac{SSE}{\chi^2_{(1-Q)/2}(n-2)}; \frac{SSE}{\chi^2_{(1+Q)/2}(n-2)}\right),$$

čia Q yra pasikliovimo lygmuo, $t_{(1-Q)/2}(n-2)$ – Stjudento skirstinio su $n-2$ laisvės laipsiais $(1-Q)/2$ lygmens kritinė reikšmė, $\chi^2_{(1-Q)/2}(n-2)$ – χ^2 skirstinio su $n-2$ laisvės laipsiais $(1-Q)/2$ lygmens kritinė reikšmė.

9.7 Hipotezė apie koeficiento b lygybę nuliui

Žinodami koeficientų įverčių skirstinius galime tikrinti hipotezes apie koeficientų reikšmes. Hipotezė $H_0 : a = 0$ aktuali tik tuo atveju, jei būtina suprasti, ar regresijos tiesė kerta koordinačių pradžios tašką. Taip būna retai. Žymiai svarbiau tikrinti hipotezę $H_0 : b = 0$, nes jei ši hipotezė pasitvirtintų, regresijos modelis virsta $Y_i = a + \varepsilon_i$, t.y. Y_i nepriklauso nuo x_i.

1. Tarkime, kad turime porinius stebėjimus $(x_1, y_1), \ldots, (x_n, y_n)$ ir tarkime, kad galioja tiesinės regresijos prielaidos.

2. Tikrinsime hipotezę:
$$\begin{cases} H_0 : b = 0, \\ H_1 : b \neq 0. \end{cases}$$

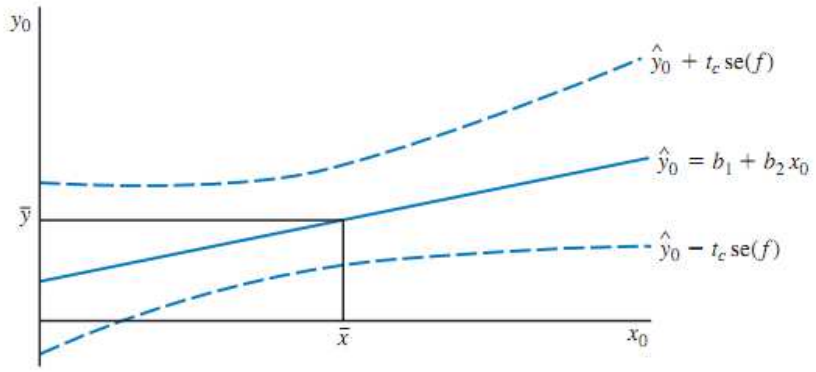
3. Kriterijaus statistika $T = \dfrac{\hat{b}}{s_b}$.

4. Tegu reikšmingumo lygmuo yra lygus α. Tuomet hipotezė H_0 atmetama, kai $|t| > t_{\alpha/2}(n-2)$, čia $t_{\alpha/2}(n-2)$ yra Stjudento skirstinio su $n-2$ laisvės laipsiais $\alpha/2$ lygmens kritinė reikšmė.

9.8 Prognozavimas regresinėje analizėje. Prognozės intervalai

Priklausomo kintamojo reikšmės prognozuojamos pagal regresijos tiesę $\hat{y}(x) = \hat{a} + \hat{b}x$, kurios koeficientai apskaičiuoti mažiausiu kvadratų metodu. Prognozė daroma konkrečiai fiksuotai nepriklausomo kintamojo reikšmei x. Tada priklausomas kintamasis
$$Y = a + bx + \varepsilon,$$
čia $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ yra atsitiktinė paklaida. Todėl tikėtina, kad nepavyks tiksliai atspėti tikrąją būsimą Y reikšmę. Kadangi ε skirstinys yra normalusis, Y taip pat yra normaliai pasiskirstęs. Normaliojo dydžio tankis didžiausią reikšmę įgyja vidurkio taške. Todėl, geriausią, ką galime padaryti prognozuoti Y reikšmę kiek galima artimesnę visų galimų (fiksuotai x reikšmei) kintamojo Y reikšmių vidurkiui. Taip ir darome, nes regresijos tiesė $\hat{y} = \hat{a} + \hat{b}x$ yra lygties $EY = a + bx$ analogas.

Viena iš dažniausių klaidų yra bandymas pernelyg tolimą ateitį spėti remiantis dabarties prielaidomis. Regresinėje analizėje prognozės daromos tik toms x reikšmėms, kurios patenka į duomenų intervalą, t.y. $\min\limits_i x_i \leq x \leq \max\limits_i x_i$. Regresijos tiesė gali gerai aprašyti kintamųjų elgesį tiriamajame intervale, kitame intervale kintamuosius gali sieti visai kita priklausomybė.



4 pav.: Prognozės intervalai

Esant fiksuotam x galima prognozuoti galimų Y reikšmių intervalą. Faktiškai sudarome skirtumo $Y - \hat{y}(x)$ prognozės intervalą. Nesunku įsitikinti, kad

$$SY = MSE\left(1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)$$

yra $Y - \hat{y}(x)$ dispersijos įvertis (*angl.* standard error of the forcast) , čia s_x^2 yra $x_1, \ldots, x_n$ empirinė dispersija. Yra žinoma, kad $(Y - \hat{y}(x))/\sqrt{SY}$ turi Stjudento skirstinį su $n-2$ laisvės laipsiais. Todėl Q pasikliovimo lygmens $y(x)$ pasikliautinasis intervalas yra

$$\hat{y}(x) \pm t_{(1-Q)/2}(n-2)\sqrt{SY}.$$

Prognozės intervalo plotis priklauso ne tik nuo imties didumo, bet ir nuo x reikšmės. Kuo x arčiau $\bar{x}$, tuo prognozės intervalas siauresnis (žr. pav. 4). Ta pati regresijos lygtis $\hat{y} = \hat{a} + \hat{b}x$ yra naudojama ir konkrečiai reikšmei $Y = a + bx + \varepsilon$ ir vidutinei reikšmei $EY = a + bx$ prognozuoti. Pirmu atveju, mes prognozuojame, pavyzdžiui, kiek konkrečiai ledų bus suvalgyta tam tikrą dieną, kai dienos temperatūra lygi 27 laipsniams, antru atveju mes prognozuojame, kiek vidutiniškai ledų bus suvalgyta dienomis, kurių vidutinė temperatūra yra 27 laipsniai. Nors pačios prognozuojamos reikšmės abiem atvejais sutampa, jų pasikliautinieji intervalai skiriasi. Antru atveju PI bus ženkliai siauresnis. Statistikos $EY - \hat{y}(x)$ dispersijos įvertis lygus

$$SEY = MSE\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right),$$

t.y. SEY nuo SY skiriasi tik vienu dėmeniu MSE. Yra žinoma, kad $(EY - \hat{y}(x))/\sqrt{SEY}$ turi Stjudento skirstinį su $n-2$ laisvės laipsiais. Todėl Q pasikliovimo lygmens $Ey(x)$ pasikliautinasis intervalas yra

$$\hat{y}(x) \pm t_{(1-Q)/2}(n-2)\sqrt{SEY}.$$