Západočeská univerzita v Plzni

Matematické modely v Ekonometrii

KMA/MME

2. semestrální práce

Autor:
Marek Lovčí

November 25, 2020

1 Zadání

Zvolte libovolný soubor dat z oblasti ekonometrického modelování. Formulujte lineární model popisující vazby dat. Odhadněte metodou nejmenších čtverců parametry modelu – odhad proveďte bodově i intervalově. Zhodnoťte kvalitu modelu. Proveďte základní regresní diagnostiku – identifikujte odlehlá a vlivná pozorování, proveďte testy residuí.

2 Data

Dataset byl vybrán z doporučené adresy `http://pages.stern.nyu.edu/~wgreene/Text/Edition7/tablelist8new.htm`. Konkrétně se jedná o set F4.1, data o prodeji Monetových obrazů. Tento dataset obsahuje 430 pozorování a 6 příznaků. Příznaky jsou cena [miliony USD] (vysvětlovaná proměnná), výška a šířka díla, binární příznak značící zda-li je dílo podepsáno (vysvětlující proměnné). Dále jsou přítomny příznaky ID obrazu (pro identifikaci opakovaných prodejů) a kód aukčního domu, kde proběhla dražba.

3 Vypracování

Definicí lineárního modelu jsme se zabývali v minulé semestrální práci. Pro vypracování této práce použijeme 3 vysvětlující proměnné (výšku, šířku obrazu a příznak podpisu) a vysvětlovanou proměnnou – prodejní cenu díla. Výsledný regresní model nebude přímkový, nýbrž nadrovinný.

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \beta_3 X_i + \epsilon_i;\ i = 1, 2, \ldots, n \tag{1}$$

Bodový odhad parametrů modelu získáme pomocí funkce *regress*.

$$\hat{\beta} = \begin{bmatrix} -4.6763 \\ 0.0902 \\ 0.1080 \\ 2.1989 \end{bmatrix} \tag{2}$$

Intervalový odhad parametrů taktéž.

$$\hat{\beta}_{int} = \begin{bmatrix} -6.3077 & -3.0449 \\ 0.0475 & 0.1330 \\ 0.0669 & 0.1490 \\ 1.2215 & 3.1764 \end{bmatrix} \tag{3}$$

Při konstrukci modelu lineární regrese nesmíme zapomenout na předpoklady, které dané konstrukci předcházejí:

1. linearita,
2. nezávislost reziduí,
3. homoskedasticita,

4. normalita reziduí.

V případě, že máme více vysvětlujících proměnných, je z pohledu statistické inference též důležitá *multikolinearita*. To je předpoklad minimální, nebo žádné lineární závislosti mezi vysvětlujícími proměnnými.

Poslední důležitou položkou, která může pomoci při zhodnocení kvality modelu jsou pákové body. Jejich počet a celkový vliv na model.

Graf 1 ukazuje vztah vysvětlované proměnné na proměnných vysvětlujících. Z grafu je možné vyčíst, že za větší částky se prodávají obrazy s autorovým podpisem. Velikost obrazu nemá vliv na cenu ve smyslu přímé úměry.

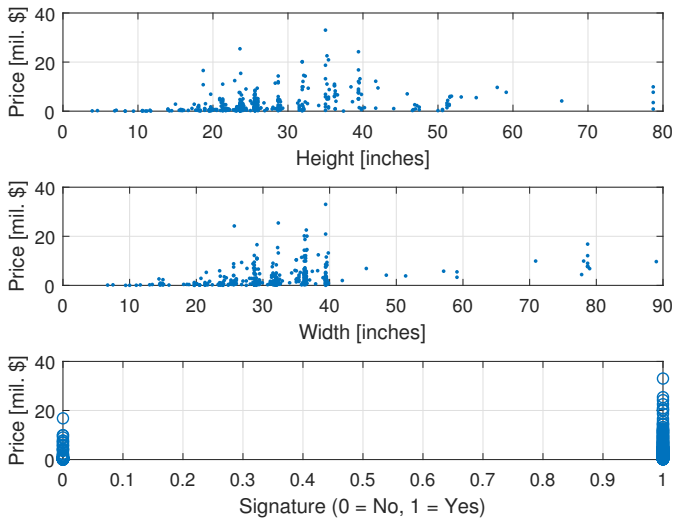

Figure 1: Vztah ceny a vysvětlujících proměnných

Tou asi nejběžnější cestou, jak zkontrolovat lineární závislost v modelu (v případě více proměnných) je vykreslení preferovaně studentizovaných reziduí proti lineárně predikované proměnné. Data jsou tímto způsobem vykreslena na grafu 2. Vizuálním testem tvrdím, že lze data aproximovat polynomem prvního stupně, data jsou tedy správně predikována lineárním modelem. Kdyby data byla rozložena ve tvaru nějaké výrazné křivky, naznačovalo by to volbu nesprávného modelu.

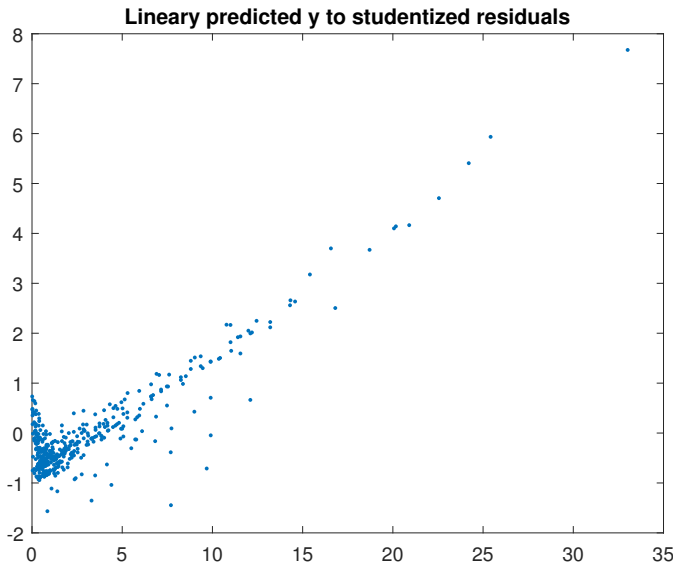

Figure 2: Zhodnocení linearity modelu

Homoskedasticitu vyhodnotíme na základě číslených hodnot v přiloženém skriptu. Pod proměnnou h se skrývá výsledek 2-výběrového F-testu shody dvou rozptylů (v tomto případě dvou částí predikované proměnné). Výsledek testu ukazuje, že rozptyly těchto částí jsou rozdílné, tedy data jsou zatížena heteroskedasticitou.

To přináší několik problémů. Heteroskedasticita může zapříčinit nesprávný odhad parametrů modelu a může ovlivnit výsledky statistik. Výsledky statistik ovlivňuje tak, že způsobuje nepřesnosti v evýpočtech p-hodnot. P-hodnoty v datech zatížených tímto problémem bývají menší než ve skutečnosti a to v důsledku způsobuje, že zkoumaný jev může být označen za statisticky signofokantní, ačkoliv ve skutečnosti není. Takže proč je v těchto datech heteroskedasticita? Domnívám se, že se jedná o podstatu dat. Výška, šířka obrazu a přítomnost autorova podpisu jsou sice zajímavými prediktory, absolutně však ignorují další faktory, jako např. historický podtext, které mohou mít vliv na finální prodejní cenu díla. Tudíž se domnívám, že přítomná heteroskedasticita je tzv. *pure* - čistá (*impure* by byla způsobená chybným modelem) a není ji třeba v mém případě více řešit.

Na následujícím grafu lze zhodnotit normalitu reziduí. Z grafu (a provedených testů normality ve skriptu) bohužel vyplývá, že rezidua nejsou normálně rozložená. Dle článků, které jsem s ohledem na tento problém pročetl, není takový stav vážným problémem. Problém bude pravděpodobně způsoben, obdobně jako u heteroskedasticity, nedostatečně reprezentativní formou vysvětlujících proměnných. Domnívám se, že by pomohlo odstraněnní heteroskedasticity a pákových bodů.

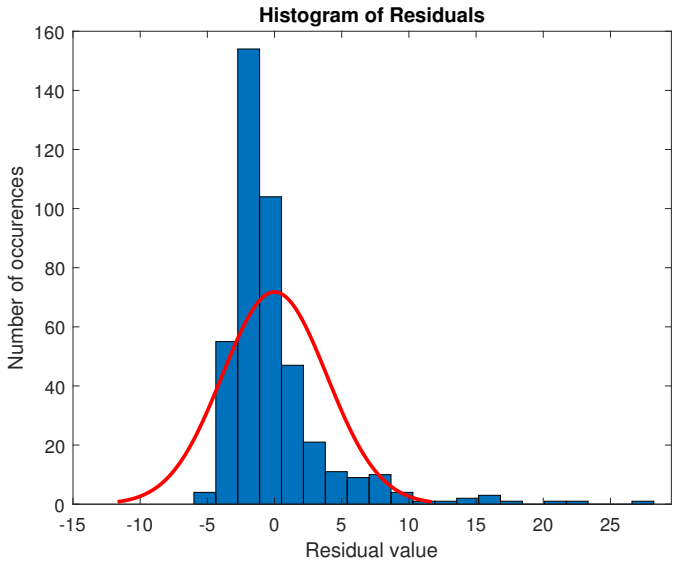

Figure 3: Normalita reziduí

Pro detekci pákových bodů byly použity 3 metriky. V závislosti na metrice bylo nalezeno 12-53 pákových bodů. 12.3 % pákových bodů je z pohledu analýzy asi celkem dost, ale zase je nutné přihlédnout k charakteru dat, jedná se o prodej uměleckých děl a lidský faktor vytváří v těchto případech hodně iregularit. Pákové body vyhodnocené Cookovou vzdáleností jsou na grafu 4.

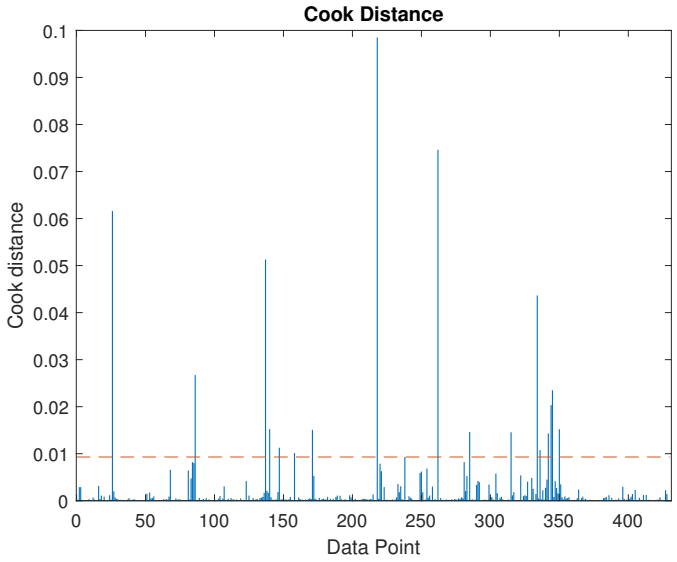

Figure 4: Pákové body dle Cooka

Na posledním grafu jsou obdobně jako na tom prvním vykresleny vztahy vysvětlujících proměnných ku (ne vysvětlované, ale...) reziduím. Z grafu samotného nic osobně nevyčtu, ale vypočtené hodnoty ukazují na velice slabou závislost, což by mělo být v souladu s předpoklady.

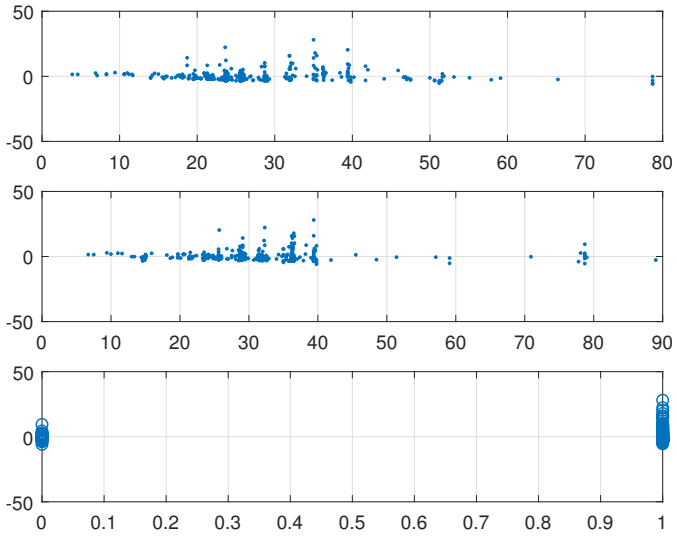

Figure 5: Závislost vysvětlujících proměnných na reziduích

4 Závěr

V semestrální práci jsme si vyzkoušeli analýzu datasetu a ověření splnění předpokladů pro tvorbu linearního modelu. Ačkoliv ne všechny výsledky vyšly dle očekávání, lze předpokládat, že hlubší analýzou podstaty dat a dodefinováním příznaků lépe vystihujících jejich pravé vlastnosti by bylo možné docílit lepších výsledků a nápravy vzniklých nesrovnalostí.