Západočeská univerzita v Plzni

Matematické modely v Ekonometrii

KMA/MME

3. semestrální práce

Autor:
Marek Lovčí

December 8, 2020

1 Zadání

Analyzujte např. šance studenta na přijetí na VŠ, kdy jsou k dispozici „vhodná“ data:

- ADMIT – binární informace o tom, zda student byl přijat na VŠ,
- GRE (Graduate Record Exam scores) – výsledky závěrečného hodnocení na SŠ,
- TOPNOTCH – binární informace o tom, zda absolvovaná SŠ patří mezi „top“ střední školy,
- GPA (Grade Point Average) – průměrné hodnocení na SŠ.

Sestavte a dokumentaci popište model charakterizující šance studenta na přijetí (pravděpodobnost) na VŠ a odhadněte parametry modelu. Parametry odhadněte několika přístupy: pomocí probitů, logitů, MNČ a doporučte model na základě vybraného kritéria kvality (např. reziduální součet čtverců). Upravte data (např. přidáním či odebráním, nebo změnou) ze souboru *data04_01.txt* nebo je možné též zvolit vlastní binární model a vlastní data.

2 Data

Pro vypracování semestrální práce nebyl zvolen avizovaný dataset o přijetí studentů, nýbrž dataset použitý již v minulé práci, tedy data o prodeji Monetových obrazů. Tento dataset obsahuje 430 pozorování a 6 příznaků. Příznaky jsou binární příznak značící zda-li je dílo podepsáno (vysvětlovaná proměnná), výška a šířka díla, cena [miliony USD] a kód aukčního domu, kde proběhla dražba (vysvětlující proměnné).

3 Vypracování

Pro zajímavější výsledky jsem výšku a šířku vynásobil, čímž jsem získal plochu obrazu a prohlásil ji za novou proměnnou. V této semestrální práci budeme tedy predikovat, zda-li je prodaný Monetův obraz podepsaný, či nikoliv, z prodejní ceny, plochy obrazu a kódu aukčního domu.

Po rychlém nahlédnutí na data zjistíme, že obsahují nerovnoměrné zastoupení vysvětlované proměnné (353 podepsaných a 77 nepodepsaných děl). Mějme na paměti, že takto biasovaná data budou (alespoň se tak domnívám) vychylovat výsledky ve prospěch toho, že dílo je podepsané.

Obrázek 1 obsahuje vykreslené vysvětlující ku vysvětlované proměnné.

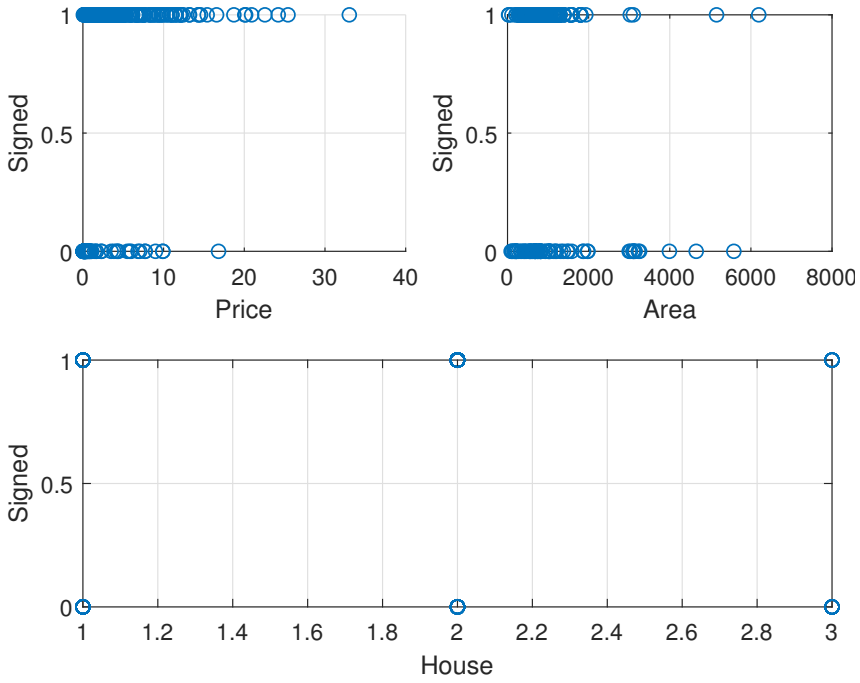

Figure 1: Vysvětlující ku vysvětlované proměnné

Nejprve složíme matici příznaků (X), zavedeme lineární regresní model a otestujeme jeho vhodnost pomocí F-testu. Po jeho vyhodnocení získáváme pro náš konkrétní model p-hodnotu rovnu 4.6336×10^{-11}. Výsledek je menší než hladina $\alpha = 5\,\%$, zamítám tedy H_0 (všechny koeficienty β jsou rovny nule), a přijímám H_A (alespoň jeden koeficien je různý od nuly). Zvolený model má tedy smysl a můžeme pokračovat dále.

Otestujeme vhodnost jednotlivých koeficientů β_i t-testem a získáváme následující p-hodnoty.

$$1.0 \times 10^{-3} \cdot \begin{bmatrix} 0.0000 \\ 0.0021 \\ 0.0000 \\ 0.5348 \end{bmatrix}$$

Vektor p-hodnot pro každý z koeficientů modelu,

Hodnoty vyšší než $\alpha = 5\,\%$ vedou k zamítnutí H_0 (koeficient je rovný nule) a značí statistickou insignifikanci daného parametru. V tomto případě kritérium nesplňuje žádný z parametrů, tudíž není třeba žádnou vysvětlující proměnnou z modelu vyřadit.

Podívejme se, jak budou vypadat modely odhadnuté pomocí *logit*, *probit* a *metodou nejmenších čtverců* při vykreslení y vůči x_i. První obrázek dává do souvislosti podpis autora s prodejní cenou obrazu. Ihned si můžeme povšimnout jevu, který je pravděpodobně způsoben rozdělením dat. I při nejnižší prodejní ceně je dle tohoto příznaku téměř 80 % pravděpodobnost, že je obraz podepsaný.

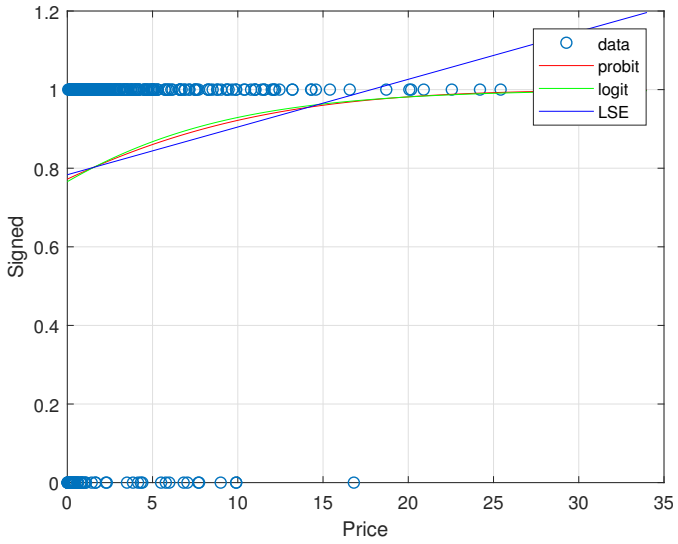

Figure 2: Srovnání modelů pro y vůči x_1

Podívejme se tedy na další graf, který dává do souvislosti podpis s plochou díla. Tento graf je zajímavý, protože regresní křivky jsou vykreslené téměř celé a pokrývají značnou část pravděpodobnostního prostoru.

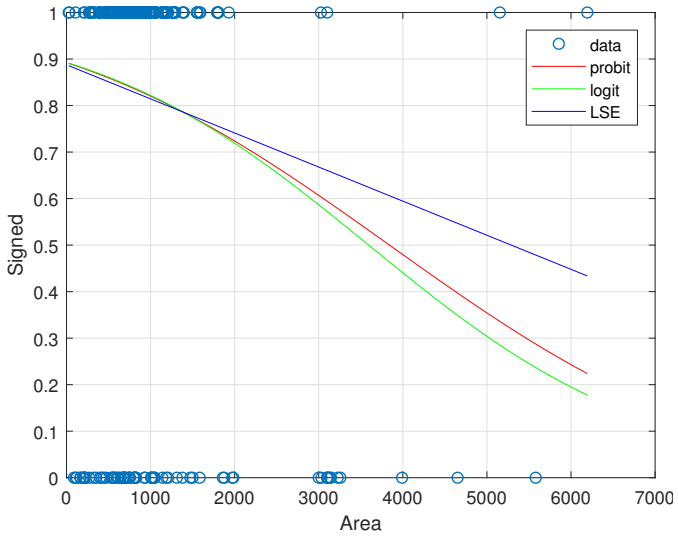

Figure 3: Srovnání modelů pro y vůči x_2

Na posledním grafu je patrný datový bias a pouze z grafu by bylo velice nepatřičné dělat jakékoliv závěry.

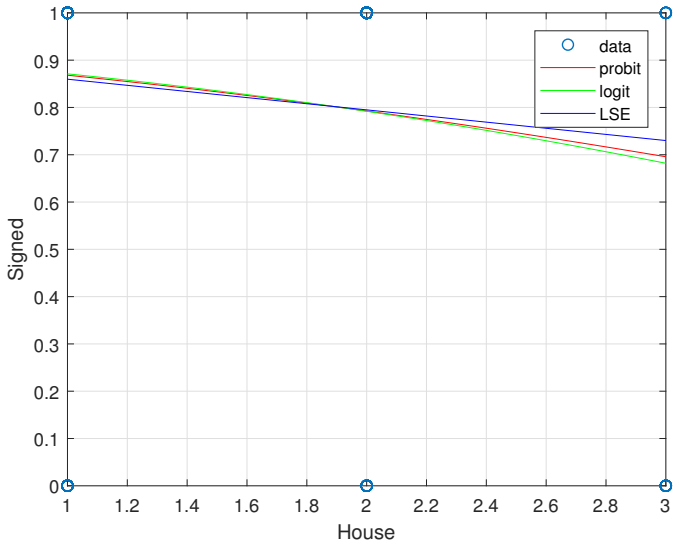

Figure 4: Srovnání modelů pro y vůči x_3

Konečnou vhodnost modelu získáme porovnáním nějaké metriky, např. *rezidálního součtu čtverců* jednotlivých modelů. Po sestavení modelů z celé matice příznaků X a vypsání reziduálního součtu pro každý z modelů získáváme následující hodnoty.

Metoda	Rezidální součet čtverců
Probit	346.3584
Logit	344.4648
MNČ	458.2582

Table 1: Rezidální součet čtverců

Z tabulky 1 přečteme, že nejmenší hodnoty jsou u metod *logit* & *probit*. Jestliže by kritériem výběru nejlepší metody bylo minimum RSČ, pak bychom za nejlepší metodu označili *logit*.

4 Závěr

V semestrální práci jsme si vyzkoušeli modelování binárního regresoru. Pomocí vizualizací jsme získali představu o fungování představených metod. Využití statistických testů nám zase upevnilo znalosti nutné pro volbu vhodných modelů a případné odstranění nevhodných proměnných.