Dynamic Pricing on Online Marketplaces: Strategies and Applications

Demand Estimation & Impact Analysis

Dr. Rainer Schlosser

Hasso Plattner Institute (EPIC)

April 19, 2016

Outline

- High Jump
- Explanatory Variables
- Goals of Today's Meeting
- Data Structure
- Least Squares Regression
- Logistic Regression
- Tips & Tricks

High Jump

- High Jump Results
- How they can be explained? What are the key factors?
- Data: Results and features of participants (observations)
- What is a suitable regression model?
- How does it work? What is the idea?
- How can we derive forecasts?
- How good are our forecasts? Is there a measure?

High Jump Data

ID	Name	Höhe	Größe	Geschlecht	Party
1	Keven	160	176	1	0
2	Martin	155	178	1	0
3	Christian	140	172	1	1
4	Matthias	150	175	1	0
5	Ralf	130	160	1	0
6	Stefan	165	190	1	1
7	Markus	165	185	1	0
8	Cindy	130	168	0	0
9	Julia	130	163	0	1
10	Anna	145	170	0	0
11	Viktoria	155	171	0	0
12	Marilena	125	167	0	0

Notations

- Number of observations?
- Which quantity do we want to explain? (dependent variable)
- Which quantities may be factors? (explanatory variables)
- What might be missing variables?
- Mean of the dependent variable? $\bar{y} = \frac{1}{N} \cdot \sum_{i=1}^{N} y_i$
- Variance of the dependent variable? $VAR = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \bar{y})^2$
- Plausibility checks: Expectations? Hypotheses?
- How do we quantify the impact/dependencies?

Least Squares Regression

- Idea: Use explanatory variables x to explain dependent variable y.
- Approach: Try to reconstruct y by linear parts of x

$$y_i \approx \beta_1 + \beta_2 \cdot x_i^{(2)} + \beta_3 \cdot x_i^{(3)} + \ldots \quad \text{where } \vec{x}_i, y_i,\ i = 1,..,N, \text{ given data}$$

β - coefficients have to be chosen such that the fit is "good".

- What is a "good" fit? We need a measure.
- Answer: Minimize the sum of squared deviations, i.e.,

$$\min_{\beta_1,\beta_2,\beta_3 \in \mathbb{R}} \sum_{i=1}^{N} \left(\beta_1 + \beta_2 \cdot x_i^{(2)} + \beta_3 \cdot x_i^{(3)} + \beta_4 \cdot x_i^{(4)} - y_i \right)^2$$

Solution & Forecasts

- We obtain optimal coefficients $\beta_1^*, \beta_2^*, \beta_3^*, \beta_4^*$ (solver/software).

- What can we do with the coefficients $\vec{\beta}^* = (-102,\ 1.43,\ 3.05,\ -5.43)$?

- (1) We can quantify the impact of factors $x^{(2)}, x^{(3)}, x^{(4)}$ on y!

- (2) We can compute smart forecasts!

- Example: We have a new participant (179 groß, männlich, Party)

- Forecast: Geschätzte Höhe = $\beta_1^* + 179 \cdot \beta_2^* + 0 \cdot \beta_3^* + 1 \cdot \beta_4^* = 151.74$

How reliable is our Model?

- We can use various combinations of explanatory variables.
- We will always obtain a result and some optimal β^* coefficients!
- How to measure the quality of a model? There is a measure: R^2 .
- Idea: How much of the variance in y can be explained by the model.
- Model fit: $\hat{y}_i = \beta_1^* + \beta_2^* \cdot x_i^{(2)} + \beta_3^* \cdot x_i^{(3)} \approx y_i$
- New variance: $VAR_{new} = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \leq VAR = \frac{1}{N} \cdot \sum_{i=1}^{N} (y_i - \overline{y})^2$
- Goodness of fit: $R^2 = 1 - \frac{VAR_{new}}{VAR} \in [0,1]$ (large is good)

Demand Estimation

Goals

- We want to measure the impact of price on sales.
- We want to estimate sales probabilities for specific scenarios.
- We just want to use observable data:

  our offer prices,

  competitors' offer prices

  our realized sales

- We want to identify the factors that determine customers' choice.

Data Structure

- We consider periods of time (e.g., days, hours, minutes)
- For every period of time we have the following raw data:

data	we	competitor 1	competitor 2	. . .
number of sales	✓	?	?	
offer price	✓	✓	✓	
product condition	✓	✓	✓	
customer rating	✓	✓	✓	
feedback count	✓	✓	✓	
shipping time	✓	✓	✓	

First Approach: Least Squares Regression

- Idea: explain „dependent variable“ by „explanatory variables“
- „dependent variable“: number of sales y (of our firm)
- „explanatory variables“: **price rank** r

  price difference to best competitor’s price

  time (day time, weekday, month etc.)

  ratings, shipping time, . . .

- Remember: Derive the β^* – coefficients for every explanatory variable by minimizing sum of squared deviations (over all observations)

First Results: Expected Sales as Function of Price

- Explanatory variable: price rank $x_i^{(2)}(a,\vec{p}) = r_i(a,\vec{p})$
- Regression result: intercept β_1^*, price rank impact β_2^*
- Expected sales: $y(a,\vec{p}) = \beta_1^* + \beta_2^* \cdot r(a,\vec{p})$ (for any situation!)
- Impact analysis: Each better rank boosts the expected number of sales by β_2^* units!

Let's be creative: Multi Linear Regression

- Invent multiple explanatory variables!
- Use transformed variables, e.g., $x^{(3)} = r^2$, $x^{(4)} = \ln(r)$, etc.
- Use and combine multiple features.
- Note, your explanatory variables can be any function of all features.
- Model: $$y(a, \vec{p}, \ldots) \approx \sum_{m=1}^{M} \beta_m \cdot x^{(m)}(a, \vec{p}, \ldots)$$

Limitations

- Advise: Do not use too many explanatory variables, i.e., $M \ll N$!

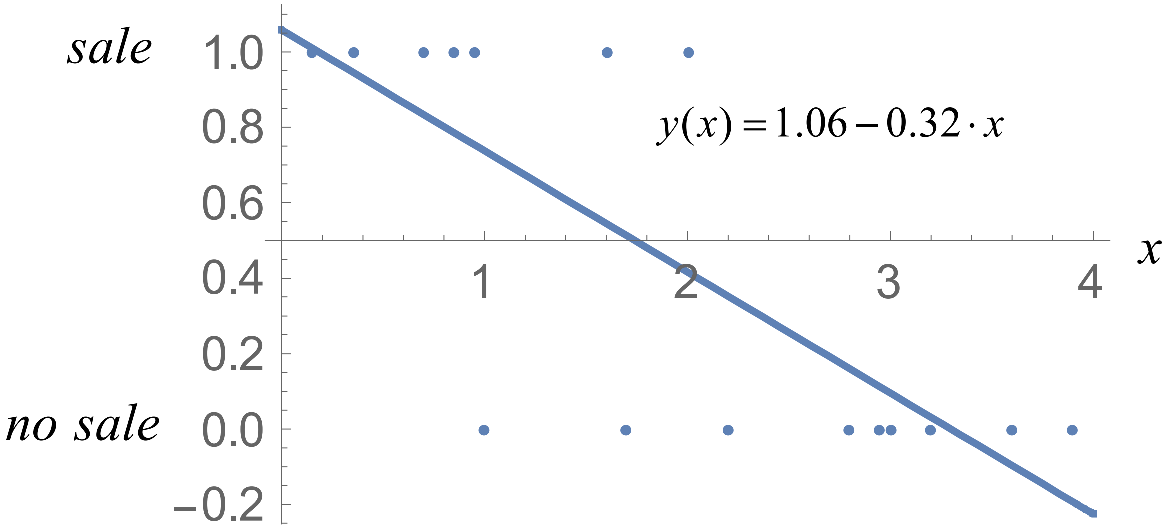
- Problems: Consider cases of rare events (short time intervals):

  # sales y is just 0 or 1, i.e., either sale or no sale.

  OLS estimations can be negative!

- Alternative: Binary Choice Models (e.g., logistic regression)

Weaknesses of Linear Regressions



Can the prediction $y(x) = 1.06 - 0.32 \cdot x$ be used as sales probability?

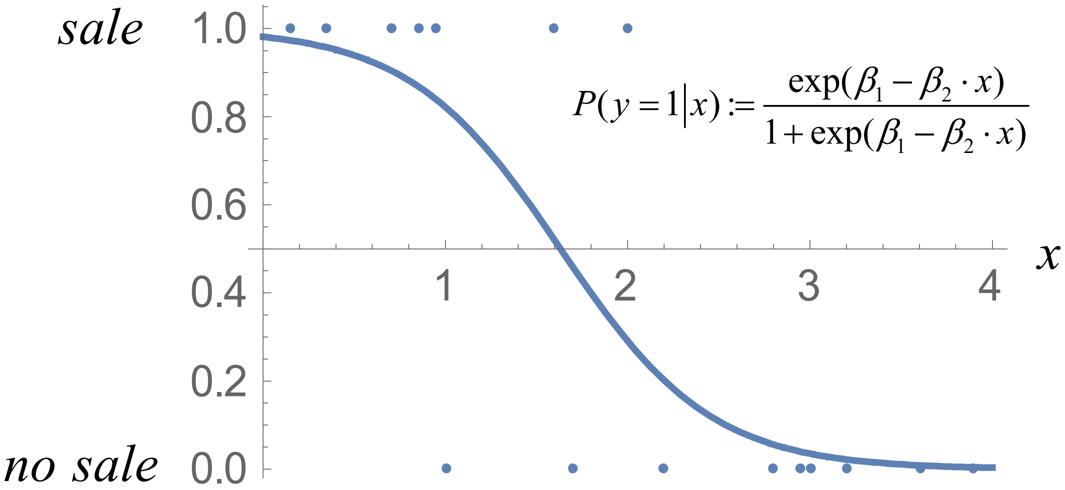
Second Approach: Logistic Regression

- Logit model: Estimate the **probability** of a binary event (sale/no sale)
- The same explanatory variables can be used
- Idea: Find a probability function that explains best sales observations by its corresponding features (explanatory variables)
- Model: $P(a, \vec{p}, \ldots) := \dfrac{\exp(L)}{1+\exp(L)} = \dfrac{1}{1+\exp(-L)}$, where

$$L(a, \vec{p}, \ldots) = \sum_{m=1}^{M} \beta_m \cdot x^{(m)}(a, \vec{p}, \ldots)$$

Illustration: Logistic Regression

- Binary 0/1 y observations, explanatory variable x, and probabilities (x)



$$P(y=1|x) := \frac{\exp(\beta_1 - \beta_2 \cdot x)}{1+\exp(\beta_1 - \beta_2 \cdot x)} \qquad \beta_1 = 3.95,\ \ \beta_1 = 2.42$$

- What are the best β coefficients? What is the Sales Probability when x=2?

Solution Details: Logistic Regression

- To solve: Find β such that the *Log-Likelihood-Function* is maximized:

$$\max_{\beta_1,\beta_2\in\mathbb{R}} \left\{ \sum_{i=1}^{N} y_i \cdot \ln p_i + (1-y_i)\cdot \ln(1-p_i) \right\}, \quad \text{where} \quad p_i = P(y_i = 1 | x_i) := \frac{e^{\beta_1+\beta_2\cdot x_i}}{1+e^{\beta_1+\beta_2\cdot x_i}}$$

- Use Standard Software (recommended)

  e.g., R, STATA, Matlab, Mathematica, ... or Libraries (Python, C, Java, ...)

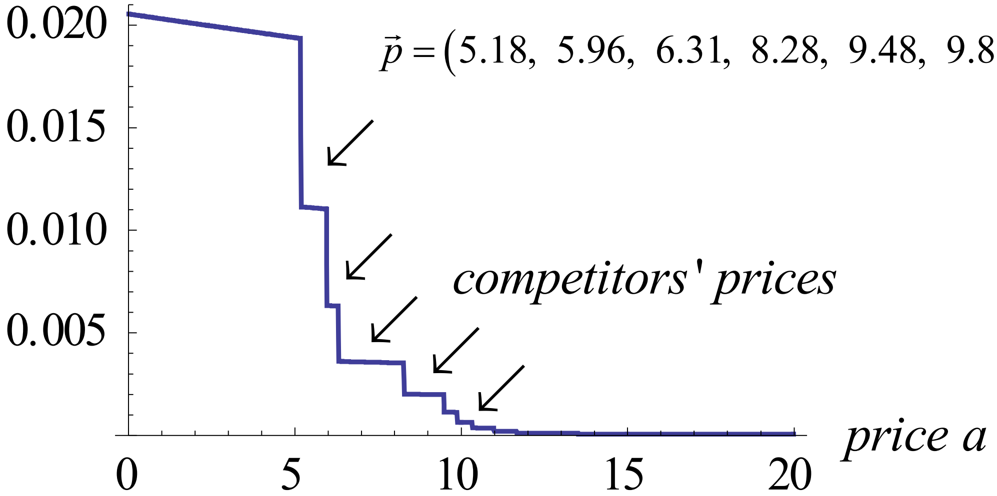
- Or: Solve an optimization problem (use nonlinear solver, e.g. AMPL/Minos)

  Optimality Conditions: $\sum_{i=1}^{N} \left(y_i - p_i(\beta_1,\beta_2)\right) \overset{!}{=} 0 \quad \& \quad \sum_{i=1}^{N} \left(y_i - p_i(\beta_1,\beta_2)\right)\cdot x_i \overset{!}{=} 0$

Result: Sales Probabilities as Function of Price

$$\textit{sales probability } p(a,\vec{p}) := \frac{e^{\beta_1+\beta_2\cdot x^{(2)}+\beta_2\cdot x^{(3)}}}{1+e^{\beta_1+\beta_2\cdot x^{(2)}+\beta_2\cdot x^{(3)}}} \qquad (\beta_1,\beta_2,\beta_3) = (\text{-}3.89, \text{-}0.56, \text{-}0.01)$$

$$\vec{p} = \left(5.18,\ 5.96,\ 6.31,\ 8.28,\ 9.48,\ 9.88,\ 10.33,\ 10.98\right)$$



Model: $x^{(2)}(a,\vec{p})$ price rank, $x^{(3)}(a,\vec{p})$ price difference to best competitor

What is a good Model?

- Compare "Goodness of fit" measures

- OLS: R^2 (high is good, share of explained variance in y)

- Logit: AIC (low is good, trade-of between fit and number of variables M)

$$AIC = -2 \cdot \sum_{i=1}^{N} \left(y_i \cdot \ln p_i + (1 - y_i) \cdot \ln(1 - p_i) \right) + 2 \cdot M$$

Note, p_i depends on all features x_i and the optimal β^* coefficients.

- Be creative: Test different variables and find the smallest AIC value.

Hint: Not quantity but quality counts!

Demand Estimation: Ideas & Tips

- Price rank
- Price difference to best competitor
- Overall price level in the competition
- Price density
- Binary variables (to be on rank *k*)
- Price differences to price rank neighbours
- Psychological Prices (e.g. does our offer price end with a “9”)

Simulate & Verify your Model

- Idea: Use random numbers to simulate features x and observed sales y

- Example: $a_i = Uniform(1,20)$, $i = 1,...,10K$ our prices for $10K$ observations

  $p_{i,c} = Uniform(1,20)$, $c = 1,...,5$ prices of 5 competitors

- Sales: $y_i = round\left(Uniform(0,1-r(a_i,\vec{p}_i)/11)\right)$, **assume** a dynamic

  or $y_i = if\ Uniform(0,1) < e^{b'x}/(1+e^{b'x})\ then\ 1\ else\ 0$ (b known)

- Check whether or not your regression finds the right coefficients

- When your model can learn different dynamics – it can learn the true one!

Overview

2	April 19	Demand Estimation
3	**April 26**	**Optimization Techniques**
4	May 3	Extensions / Projects A-D
5	May 10	Assign Projects to Teams
6	May 17	no Meeting
7	May 24	Workshop / Group Meetings
8	May 31	Workshop / Group Meetings
9	June 7	Presentations (First Results)
10	June 14	Workshop / Group Meetings
11	June 21	Workshop / Group Meetings
12	June 28	Workshop / Group Meetings
13	July 5	no Meeting
14	**July 12**	**Presentations (Final Results), Discussions, Feedback**