Robust and quantile regression

Outline

Managerial motivation

The model

Case study: the bakery history

The dual reading: the support points

Quantile regression and safety stock

Feature selection with a budget

Exercises

Learning objectives (1/2)

At the end of this chapter you will be able to:

- formulate regression with absolute deviations as an LP;
- read its dual, recognising the points that support the estimate and the quantile property;

Learning objectives (2/2)

At the end of this chapter you will be able to:

- estimate the quantity to produce conditional on the conditions of the day;
- select the relevant features with a budget on the coefficients, reading its shadow price.

Prerequisites

- From this module, Chapter 7 «The Newsvendor and its variants»: the quantile rule and the critical fractile
- From this module, Chapter 2 «Background: linear programming and sensitivity analysis»: the dual of an LP and shadow prices

Managerial motivation

Introduction: Managerial motivation

- Every model seen so far starts from known parameters: the costs of production, the hourly prices of charging, the scenarios of the newsvendor
- Somebody, however, has to estimate those numbers
- The estimation is itself an optimization problem, solved with the same solver as the rest of the course

Predicting a quantity

- Here we use optimization to *predict a quantity*
- The choice of the objective function becomes a managerial decision:
  - minimising the sum of the *absolute* deviations keeps the problem an LP
  - the estimate is not dragged around by anomalous days
  - weighting deviations above and below asymmetrically aims at a *quantile* of demand instead of its centre
- That is exactly the number the newsvendor needs

Summary

- The parameters of the models have to be estimated too, and the estimation is an optimization problem
- With absolute deviations, regression is an LP
- Asymmetric weights on the deviations give a quantile: the number the newsvendor needs

The model

Introduction: The model

- We formulate quantile regression as an LP
- We see how to linearise the absolute value with two non-negative variables
- An example by hand shows that, with no features, the model returns the empirical quantile

The problem in words

- *We decide* the coefficients of a linear function that, from the observed features of a day, predicts demand
- *The objective* is to make the deviations between observed and predicted demand small, weighting those above and those below asymmetrically
- *The constraints* define, observation by observation, the deviation of the prediction from the datum

Data (model input) (1/2)

Symbol	Type	Meaning
n	$\in \mathbb{Z}_{\geq 1}$	number of historical observations, indexed by $i \in \{1, 2, \ldots, n\}$
p	$\in \mathbb{Z}_{\geq 0}$	number of observed features, indexed by $j \in \{1, 2, \ldots, p\}$
x_{ij}	$\in \mathbb{Q}$	value of feature j in observation i

Data (model input) (2/2)

Symbol	Type	Meaning
y_i	$\in \mathbb{Q}$	observed value to be explained in observation i (demand)
τ	$\in \mathbb{Q}$, $0 < \tau < 1$	quantile aimed at: $\tau = 1/2$ gives the median, a large τ a prudent prediction on the high side

Decision variables (1/2)

We introduce $p + 1$ free variables for the function to be estimated:

$$\begin{cases} w_j = \text{coefficient of feature } j \\ b = \text{intercept of the function} \end{cases} \qquad \forall j \in \{1, 2, \ldots, p\}$$

Decision variables (2/2)

For each observation, two non-negative variables measuring the deviation of the prediction from the datum:

$$\begin{cases} u_i = \text{deviation above of observation } i \\ v_i = \text{deviation below of observation } i \end{cases}$$

$$\forall i \in \{1, 2, \ldots, n\}$$

LP model of quantile regression

LP model of quantile regression

$$\min \quad \sum_{i=1}^{n} \left[\tau\, u_i + (1-\tau)\, v_i \right] \tag{10.1a}$$

$$\text{subject to} \quad \sum_{j=1}^{p} w_j\, x_{ij} + b + u_i - v_i = y_i, \quad \forall i \in \{1, 2, \ldots, n\}, \tag{10.1b}$$

$$w_j \gtreqless 0, \quad \forall j \in \{1, 2, \ldots, p\}, \tag{10.1c}$$

$$b \gtreqless 0, \tag{10.1d}$$

$$u_i,\ v_i \geq 0, \quad \forall i \in \{1, 2, \ldots, n\}. \tag{10.1e}$$

Description of the LP model

- The linear objective function pays τ for every unit of deviation above and $1 - \tau$ for every unit of deviation below:
  - with $\tau = 1/2$ it is half the sum of the absolute deviations
  - with a large τ falling short of the datum costs more than overshooting it
- The linear deviation constraints define the deviation of every observation as the difference of its two parts, $u_i - v_i = y_i -$ prediction (n linear constraints)
- The last constraints define the variables: free coefficients and intercept, non-negative deviations

Why two variables for a single deviation (1/2)

- The deviation $y_i - \left(\sum_j w_j x_{ij} + b\right)$ has an arbitrary sign, but the objective must pay its *absolute value*
- Splitting it into the positive part u_i and the negative part v_i, both non-negative, keeps the objective linear

Why two variables for a single deviation (2/2)

- At the optimum at least one of the two is zero
- If both were positive, reducing them by the same amount would leave the deviation constraint satisfied and would decrease the objective, because $\tau > 0$ and $1 - \tau > 0$
- It is the same positive-part trick already used for the scenario newsvendor

Example: With no features, the empirical quantile comes back (1/4)

- With $p = 0$ only the intercept is left and the model chooses a single number b
- Seven days of observed sales:

$$88,\ 96,\ 104,\ 112,\ 120,\ 132,\ 148 \text{ pieces}$$

- The bakery of the newsvendor has $c_u = 9$ and $c_o = 4$, hence

$$\tau = \alpha^* = 9/13 = 0.6923$$

Example: With no features, the empirical quantile comes back (2/4)

Step 1 — what it costs to move b

- Moving b up by one unit saves τ on every observation lying *above* b
- And costs $1 - \tau$ more on every observation lying *below*
- It pays to go up as long as the observations above are numerous enough to compensate

Example: With no features, the empirical quantile comes back (3/4)

Step 2 — the optimum

- The optimal value is $\tilde{b} = 120$, the fifth observation in increasing order
- 4 lie below it and 2 above, and

$$4/7 = 0.5714 \leq \tau \leq 5/7 = 0.7143$$

- The optimal cost is

$$\tilde{z} = \frac{4}{13}(32 + 24 + 16 + 8) + \frac{9}{13}(12 + 28)$$
$$= \frac{320+360}{13} = \frac{680}{13} = 52.3077$$

Example: With no features, the empirical quantile comes back (4/4)

Step 3 — the duals

$$\tilde{\boldsymbol{\pi}} = \left(-\tfrac{4}{13}, -\tfrac{4}{13}, -\tfrac{4}{13}, -\tfrac{4}{13}, -\tfrac{2}{13}, \tfrac{9}{13}, \tfrac{9}{13}\right)$$

- $\tau = 9/13$ on the observations above the estimate, $-(1-\tau) = -4/13$ on those below, an interior value on the interpolated one
- Their sum is zero and the dual value $\sum_i \tilde{\pi}_i\, y_i = 680/13$ matches the primal one

With no features the model returns the **empirical quantile**: the quantile rule of the newsvendor, written as an LP. With features it becomes a quantile *conditional* on the conditions of the day.

Summary

- Quantile regression is an LP: deviation $= u_i - v_i$, with $u_i, v_i \geq 0$
- The objective pays τ for the deviations above and $1 - \tau$ for those below
- At the optimum at least one of u_i and v_i is zero
- With no features we obtain the empirical quantile of level τ

Case study: the bakery history

Introduction: The bakery history

- We apply the model to sixty days of sales of the bakery
- We estimate the line twice: on all the days and on the normal days only
- We will see that the LP withstands the anomalous days

The data

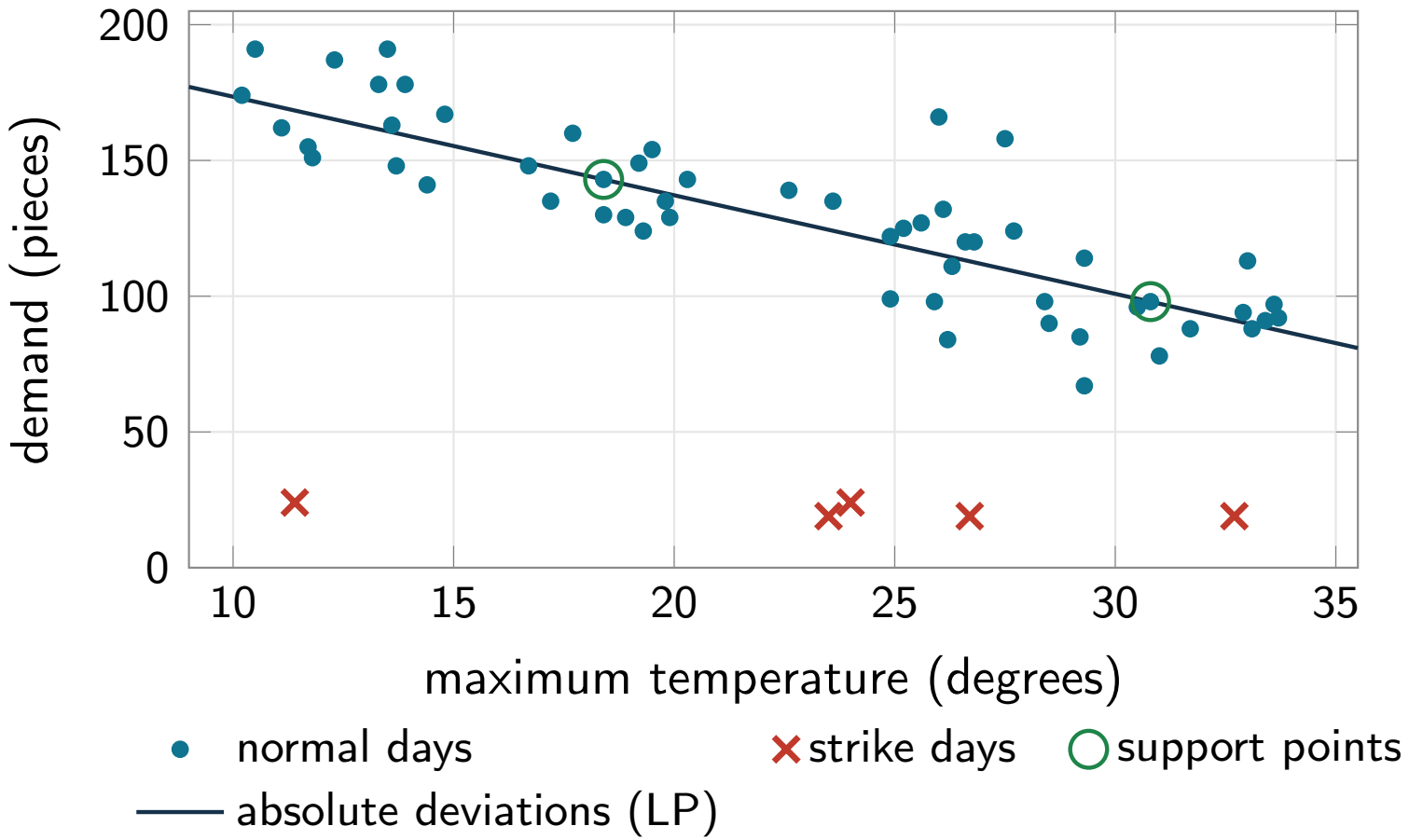
- Sixty days of sales of the bakery: for each day the demand in pieces, the maximum temperature, the price charged and a weekend indicator
- Five of those days fall during two transport strikes: the town centre is deserted and sales collapse to about twenty pieces
- We first estimate the relation with temperature alone ($p = 1$), which can be drawn in a chart

Absolute deviations with and without the strike days

model ($\tau = 1/2$)	estimated line
LP on the 60 days	demand $= 209.77 - 3.63 \cdot$ temperature
LP without the 5 strike days	demand $= 210.49 - 3.58 \cdot$ temperature

- Mean absolute deviation over the 55 normal days: 13.08 pieces
- Distance between the two intercepts: a mere 0.72 pieces

Five strike days do not move the LP line



Why the LP withstands anomalous data

- The LP pays every deviation *once*, in proportion to its size
- The line stays where the normal data are: the model reaches on its own the answer one would get by cleaning the history by hand
- The two points circled in green are the support points, the only ones the line interpolates
- Minimising the *squared* deviations leads to least squares, a quadratic model: in the module *NLP Modelling*, "Least-squares regression"

Summary

- Absolute deviations (LP): the line stays where the normal data are
- With and without the strike days the intercepts differ by 0.72 pieces
- A model that needs no supervision: it can be left running every week

The dual reading: the support points

Introduction: The dual reading

- We associate a dual with every deviation constraint
- Complementary slackness says which observations determine the estimate: the **support points**
- The dual of the intercept is the quantile property

The window of the duals

Associating a dual π_i with each deviation constraint:

- optimality on the non-negative variables u_i and v_i confines every dual to a window that depends on τ alone:

$$-(1-\tau) \;\leq\; \tilde{\pi}_i \;\leq\; \tau, \qquad \forall i \in \{1, 2, \ldots, n\}$$

- optimality on the free variables w_j and b gives the two conditions

$$\sum_{i=1}^{n} \tilde{\pi}_i \, x_{ij} = 0 \quad \forall j \in \{1, 2, \ldots, p\}, \qquad \sum_{i=1}^{n} \tilde{\pi}_i = 0$$

Complementary slackness

Complementary slackness says where each dual sits:

Position of the observation	Positive variable	Dual
above the estimate (y_i larger)	$\tilde{u}_i > 0$	$\tilde{\pi}_i = \tau$
below the estimate (y_i smaller)	$\tilde{v}_i > 0$	$\tilde{\pi}_i = -(1-\tau)$
on the estimate (interpolated)	$\tilde{u}_i = \tilde{v}_i = 0$	$-(1-\tau) < \tilde{\pi}_i < \tau$

The support points

- The interpolated observations are the **support points**
- They are the only ones whose dual is not stuck at an endpoint, and they are what determines the estimate
- In general there are $p + 1$ of them, as many as the free variables of the model
- Moving any of the other observations, as long as it stays on its side of the estimate, changes nothing

The duals on the bakery history

Line with temperature alone, $\tau = 1/2$:

observations above the line ($\pi = +0.5$)	29
observations below the line ($\pi = -0.5$)	29
interpolated observations ($= p + 1$)	2
sum of the duals	0.000000
sum of the duals times temperature	0.000000

Dual value 616.9960 = primal value 616.9960 (strong duality)

The dual of the intercept is the quantile property (1/2)

- The condition $\sum_i \tilde{\pi}_i = 0$ is the dual of the intercept
- It is the quantile property written algebraically: substituting the values of the table

$$\tau\, n_+ - (1-\tau)\, n_- + \sum_{i \text{ interpolated}} \tilde{\pi}_i = 0$$

where n_+ and n_- count the observations above and below the estimate

The dual of the intercept is the quantile property (2/2)

- Up to the $p+1$ support points, the estimate therefore leaves a fraction τ of the observations below it
- With $\tau = 1/2$, half above and half below (29 and 29 in the example)
- The condition $\sum_i \tilde{\pi}_i x_{ij} = 0$ says the same thing feature by feature: the sign of the deviations must retain no link with the data

Few points sustain the estimate

- Very few observations, the support points, sustain the solution
- The others could disappear without changing it, as long as they stay on their side of the estimate
- It is the same complementarity seen in the background chapter

The deviation as a diagnostic tool

- We estimate the model with all three features on the 60 days:

$$\text{demand} = 276.46 - 3.97\,\text{temp.} - 23.12\,\text{price} + 28.51\,\text{weekend}$$

- Absolute deviation: at most 23.2 pieces on normal days, at least 67.5 on the five strike days
- A threshold at 40 pieces isolates exactly days $12, 23, 30, 42, 47$: they are the strike days
- The model finds them knowing nothing about the strikes

Summary

- Every dual lives in the window $[-(1-\tau), \tau]$
- The interpolated observations ($p+1$ in general) are the support points and determine the estimate
- $\sum_i \tilde{\pi}_i = 0$ is the quantile property
- The deviation of the robust model identifies the anomalous days

Quantile regression and safety stock

Introduction: Quantile regression

- Changing τ estimates a different quantile of demand
- With τ equal to the critical fractile of the newsvendor we obtain directly the quantity to produce
- The difference from the median is the safety stock

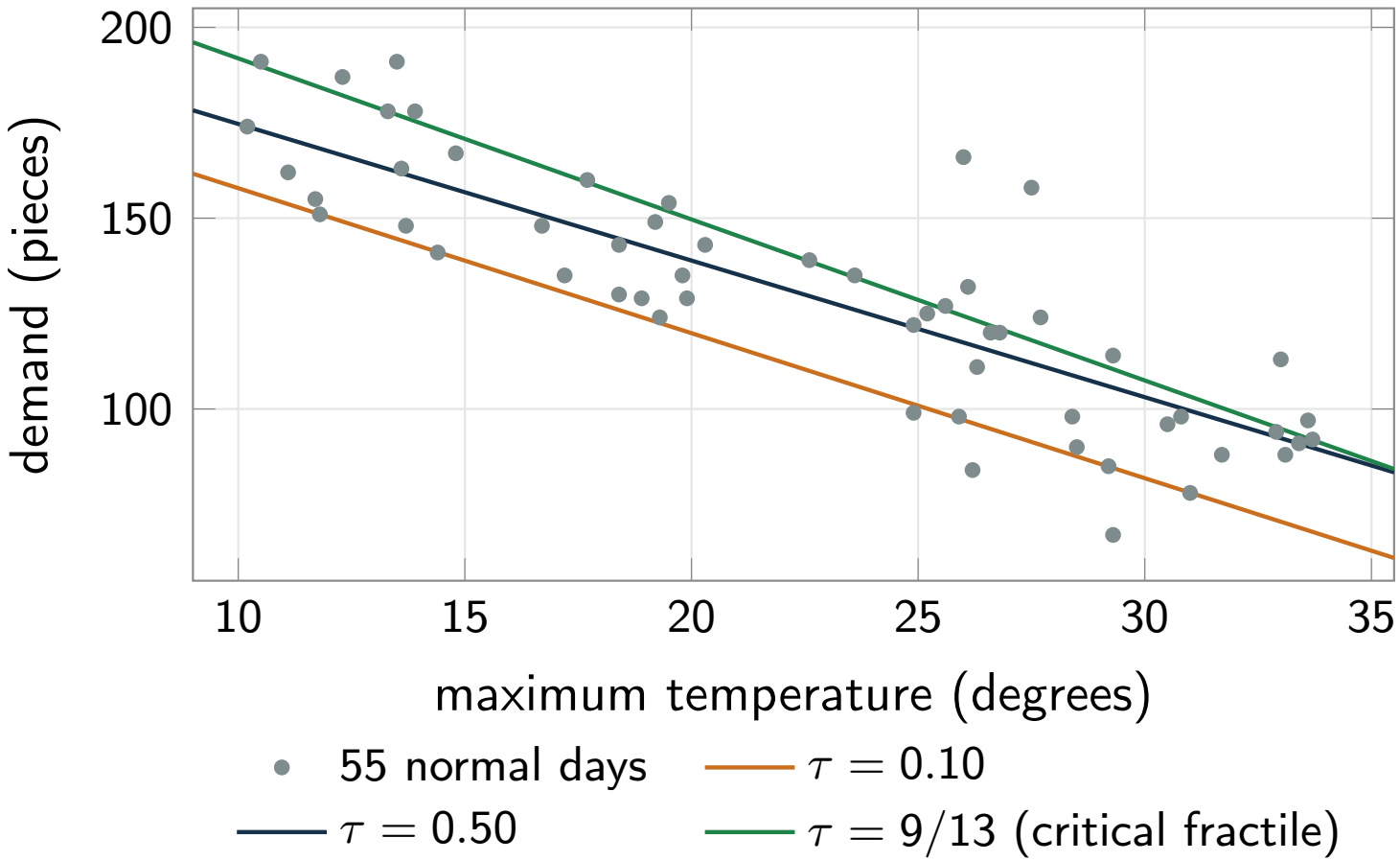
The three-feature model for several τ

On the 55 normal days; the last column checks the quantile property.

τ	intercept	temp.	price	weekend	share below
0.10	269.29	-4.04	-24.71	$+37.05$	0.073
0.50	272.01	-3.85	-22.24	$+28.74$	0.455
9/13	281.19	-3.96	-23.21	$+31.47$	0.673
0.90	287.49	-3.57	-25.69	$+27.98$	0.873

- $9/13 = 0.6923$ is the critical fractile of the bakery
- "share below": fraction of observations falling below the estimate

One line per quantile: the fan of demand



How to read the chart

- The lines estimated with temperature alone for three values of τ
- Each leaves the corresponding fraction of observations below it
- The green line, with τ equal to the critical fractile of the bakery, is the quantity to bake day by day

A hot weekend day at full price

28 degrees, price 3.00: the model gives two different numbers depending on τ.

median predicted demand	126.3 pieces
predicted 0.6923 quantile (quantity to bake)	132.1 pieces
safety stock	5.8 pieces
0.6923 quantile with no features	148.0 pieces

The conditional quantile (1/2)

- The quantile rule of the newsvendor still holds, but the number it applies to changes
- The quantile computed over the whole history says to bake 148 pieces
- The one *conditional* on the conditions of the day asks for 132
- Sixteen unsold pieces a day fewer, without touching the costs or the decision rule: just by using information already in hand

The conditional quantile (2/2)

- The gap between the median estimate and the one at the critical fractile, 5.8 pieces, is the **safety stock**
- The model computes it directly, with no assumption about the distribution

Summary

- Every τ gives a line that leaves roughly a fraction τ of the observations below it
- With τ = critical fractile we obtain the quantity to produce, conditional on the conditions of the day
- Safety stock = quantile at the critical fractile − median, with no assumption about the distribution

Feature selection with a budget

Introduction: Feature selection

- Adding features always improves the fit on the historical data, even when they are irrelevant
- Given enough freedom the model ends up chasing noise
- One way to govern this is to put a *budget* on the coefficients

New variables and a new datum

We need p new non-negative variables:

$$z_j = \text{coefficient budget assigned to feature } j$$

$$\forall j \in \{1, 2, \ldots, p\}$$

and the datum $t \in \mathbb{Q}_{\geq 0}$, the total budget.

The budget constraints

The LP model of quantile regression gains the constraints

$$-z_j \leq w_j \leq z_j, \qquad \forall j \in \{1, 2, \ldots, p\}$$

$$\sum_{j=1}^{p} z_j \leq t$$

$$z_j \geq 0, \qquad \forall j \in \{1, 2, \ldots, p\}$$

- The first force z_j to be at least $|w_j|$ (at the optimum it equals it exactly, because z_j consumes budget)
- The constraint $\sum_j z_j \leq t$ shares the budget out among the features
- The model remains an LP

Standardising the features

Observation 1: Different units of measurement

The budget adds up coefficients of features measured in different units: degrees, euros, 0/1 indicators. Without standardising, the constraint $\sum_j z_j \leq t$ would compare incommensurable quantities and the selection would depend on the units of measurement.

Here every feature is brought to zero mean and unit standard deviation, so each coefficient reads as pieces per standard deviation of the feature.

Four irrelevant features

- To the three real features we add four with no link whatsoever to demand:
  - social followers
  - rainfall in a neighbouring region
  - a stock index
  - the day of the month
- As the budget grows we observe which ones enter

Coefficients as the budget grows (1/2)

Real features:

budget	temperature	price	weekend
10	−10.00	0.00	0.00
20	−18.16	−1.51	0.00
30	−22.43	−4.45	2.35
45	−26.98	−8.22	9.09

Coefficients as the budget grows (2/2)

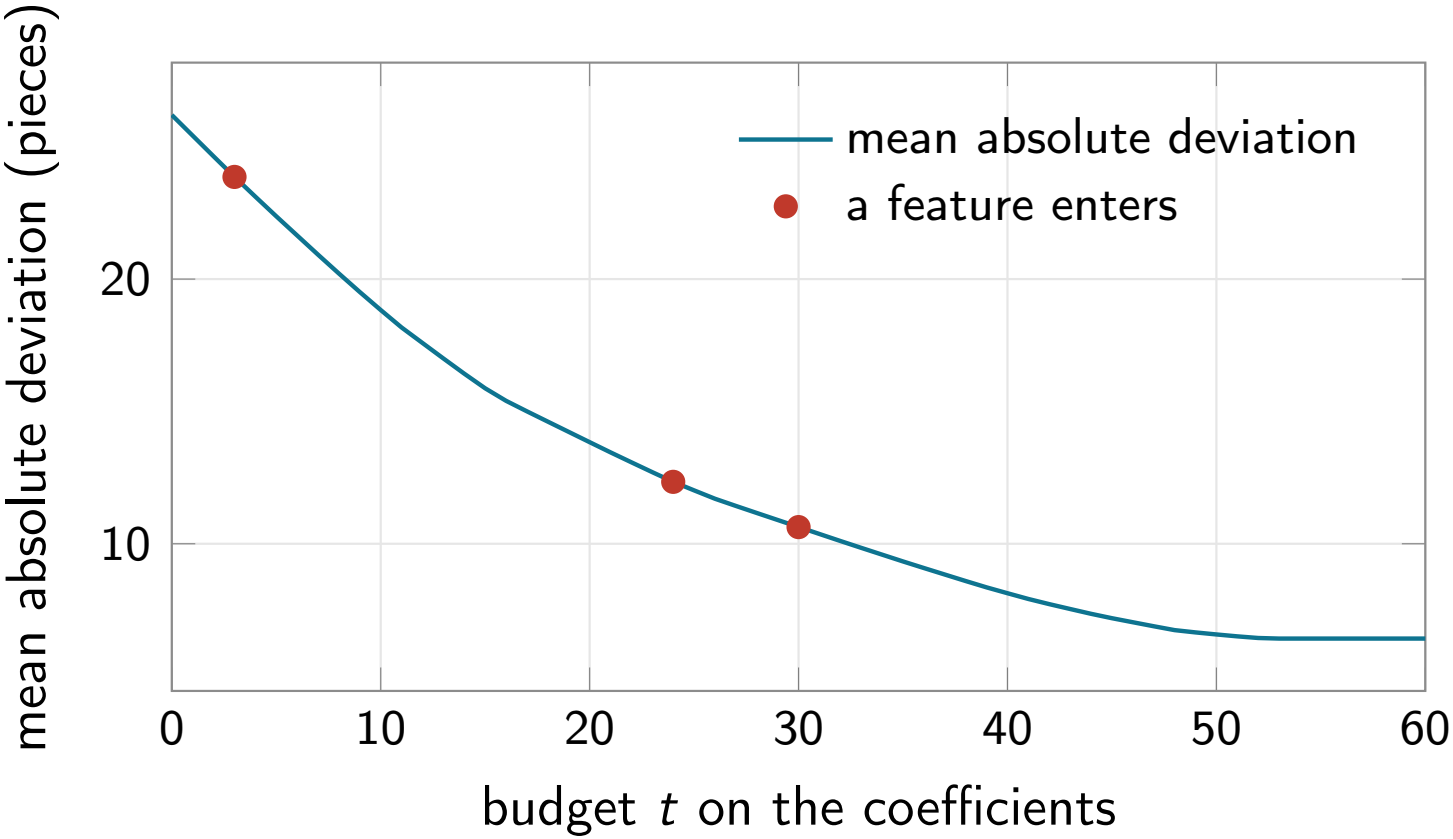
Irrelevant features:

budget	followers	rain	index	day of the month
10	0.00	0.00	0.00	0.00
20	0.00	−0.33	0.00	0.00
30	0.00	−0.77	0.00	0.00
45	0.00	−0.63	0.00	−0.09

Mean deviation and shadow price of the budget

budget	mean deviation (pieces)	shadow price (pieces per unit)
5	22.37	−19.952
10	18.83	−18.665
20	13.84	−10.658
30	10.63	−7.271
45	7.19	−4.197

The price of complexity: error against budget



How to read the chart

- The mean absolute deviation as a function of the budget on the coefficients
- A convex decreasing curve, like the optimal value of an LP as the right-hand side varies
- The red dots mark the budget at which temperature ($t = 3$), price ($t = 24$) and weekend ($t = 30$) enter, in this order

The price of complexity (1/2)

- The shadow price of the constraint $\sum_j z_j \leq t$ is the **price of complexity**: how many pieces of error are saved for every extra unit of budget
- It is -19.95 pieces per unit when the budget is scarce and -4.20 when it is plentiful
- The first units buy temperature, which explains almost everything; the last ones buy noise

The price of complexity (2/2)

- The three real features enter in the order of their importance: temperature, price, weekend
- None of the four irrelevant ones exceeds 2 pieces per standard deviation until the budget becomes plainly excessive
- The budget to choose is not the one minimising the historical error, but the one beyond which the shadow price becomes negligible

Summary

- A budget t on the sum of the $|w_j|$ selects the features, while remaining an LP
- The features must be standardised first
- The shadow price of the budget is the price of complexity
- One chooses the budget beyond which the shadow price becomes negligible

Exercises

Introduction: Exercises

- The exercises test the concepts of the chapter
- They are solved by hand or with the lab script and Gurobi
- They touch the quantile by hand, the support points, asymmetric costs, the convexity of the budget and validation

Exercise: quantile by hand

- With the seven days of the example

$$88,\ 96,\ 104,\ 112,\ 120,\ 132,\ 148$$

  and $\tau = 1/2$, compute by hand the median, the optimal value of the objective and the seven duals

- Check with the script that the duals sum to zero

Exercise: support points

- Estimate the model with temperature alone and $\tau = 1/2$
- Move an observation lying above the line up by $+30$ pieces and re-solve: does the estimate change?
- And if one of the two support points is moved down by -30 pieces?
- Explain through complementary slackness

Exercise: asymmetric cost of error

- The bakery considers selling the leftovers off at the end of the day, taking c_o from 4 to 2 €
- Recompute α^* and re-run the quantile regression
- By how many pieces does the production of the typical day change?

Exercise: budget and convexity

- Plot the optimal value of the LP with a budget for $t \in \{0, 1, \ldots, 60\}$
- Check that it is convex and piecewise linear
- Locate the breakpoints and show that they coincide with the budgets at which the shadow price changes value

Exercise: validation

- Split the 55 normal days into estimation (40) and validation (15)
- For $t \in \{5, 10, \ldots, 55\}$ plot the mean absolute deviation on the two sets
- Is the estimation one monotone in t? And the validation one?
- Which budget do you choose?

Final summary (1/2)

The model:

- Quantile regression as an LP: $u_i, v_i \geq 0$ paid τ and $1 - \tau$
- With no features: the empirical quantile

Robustness:

- Absolute deviations ignore the anomalous days
- The deviation of the robust model identifies the anomalies

Final summary (2/2)

Dual reading:

- Duals in the window $[-(1-\tau), \tau]$; $p+1$ support points
- $\sum_i \tilde{\pi}_i = 0$ is the quantile property

Decisions:

- $\tau =$ critical fractile: quantity to produce and safety stock
- Budget on the coefficients: the shadow price is the price of complexity