SH Dr. Soufiane Hourri

drsoufiane.com

DUT · SEMESTER 3 · COMPUTER SCIENCE

Machine Learning

Practice 02 · Linear regression in practice

Dr. Soufiane Hourri, PhD, HDR

Associate Professor & AI Consultant

École Supérieure de Technologie de Safi · Cadi Ayyad University

2026-2027

Today's practice

A Body mass from flipper length (guided)

B Adding categories: species and sex

C Restaurant tips (on your own)

D Bonus: tip percentage and extrapolation

TODAY'S GOAL

Run the whole regression workflow yourself, and explain every number you print.

Learning objectives

By the end of this practice, you will be able to

1. Split, fit and evaluate a linear regression in scikit-learn
2. Compare a model with the mean baseline using MAE, RMSE and R^2
3. Interpret a slope in the units of the data
4. Explain why adding a category can change a slope
5. Read a residual plot and spot extrapolation

YOU NEED

- JupyterLab (Practice 01)
- Practice_02_Notebook.ipynb
- penguins.csv and tips.csv in the same folder

PRACTICE 02 · REMINDER

The workflow, step by step

- Read the data and remove the rows you cannot use
- Split: 80% training, 20% test, random_state=42
- Baseline: DummyRegressor predicts the training mean
- Fit LinearRegression on the training set only
- Evaluate on the test set: MAE, RMSE, R^2
- Interpret the coefficients, with units

KEY TAKEAWAY

A model is useful only if it beats the baseline on the test set.

PRACTICE 02 · REMINDER

The three scores

$$\mathrm{MAE} = \mathrm{mean}\ |y - \hat{y}| \qquad \mathrm{RMSE} = \sqrt{(\ \mathrm{mean}\ (y - \hat{y})^2\ )}$$

$$R^2 = 1 - \Sigma(y - \hat{y})^2 \ /\ \Sigma(y - \bar{y})^2$$

SYMBOLS

y : true value; $\hat{y}$: prediction

$\bar{y}$: mean of the true values

MAE and RMSE: in the unit of y

READING

Lower MAE and RMSE are better.

R² near 1 is good; R² ≤ 0 means no better than the mean.

PRACTICE 02 · SETUP

Getting started

CODE · PYTHON

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.dummy import DummyRegressor
from sklearn.metrics import (
    mean_absolute_error, root_mean_squared_error,
    r2_score)

df = pd.read_csv("penguins.csv").dropna()
print(df.shape)
```

OUTPUT

```
(333, 7)
```

344 penguins, 11 with a missing value: 333 remain.

A

PART

Body mass from flipper length

Can we predict how heavy a penguin is from the length of its flipper?

PART A · GUIDED CODE

Part A · Split and baseline

CODE · PYTHON

```python
train, test = train_test_split(df, test_size=0.2,
                               random_state=42)
X_tr = train[["flipper_length_mm"]]
X_te = test[["flipper_length_mm"]]
y_tr = train["body_mass_g"]
y_te = test["body_mass_g"]

base = DummyRegressor().fit(X_tr, y_tr)
p = base.predict(X_te)
print(len(train), len(test), p[0])
print(mean_absolute_error(y_te, p),
      r2_score(y_te, p))
```

OUTPUT

```
266 67 4214.8
674.1 -0.002
```

Rounded. The baseline predicts the training mean for everyone.

PART A · GUIDED CODE

Part A · Fit and evaluate

CODE · PYTHON

```python
lr = LinearRegression().fit(X_tr, y_tr)
print(lr.coef_, lr.intercept_)

p = lr.predict(X_te)
print(mean_absolute_error(y_te, p))
print(root_mean_squared_error(y_te, p))
print(r2_score(y_te, p))
```

OUTPUT

```
[50.4] -5919.3
290.1
360.4
0.794
```

Rounded.

PART A · EXERCISE

Part A · Questions

GOAL

Interpret the model of the previous slide, in grams.

TIME

15 min

QUESTIONS

1. Interpret the slope for 1 mm and for 10 mm
2. Predict the mass for 200 mm and 150 mm; trust both?
3. Compare MAE and R^2 with the baseline
4. Plot the test residuals (y_te − p) against flipper length

DELIVERABLE

One sentence per question, with units.

PART A · CHECK

Part A · What your residual plot should look like

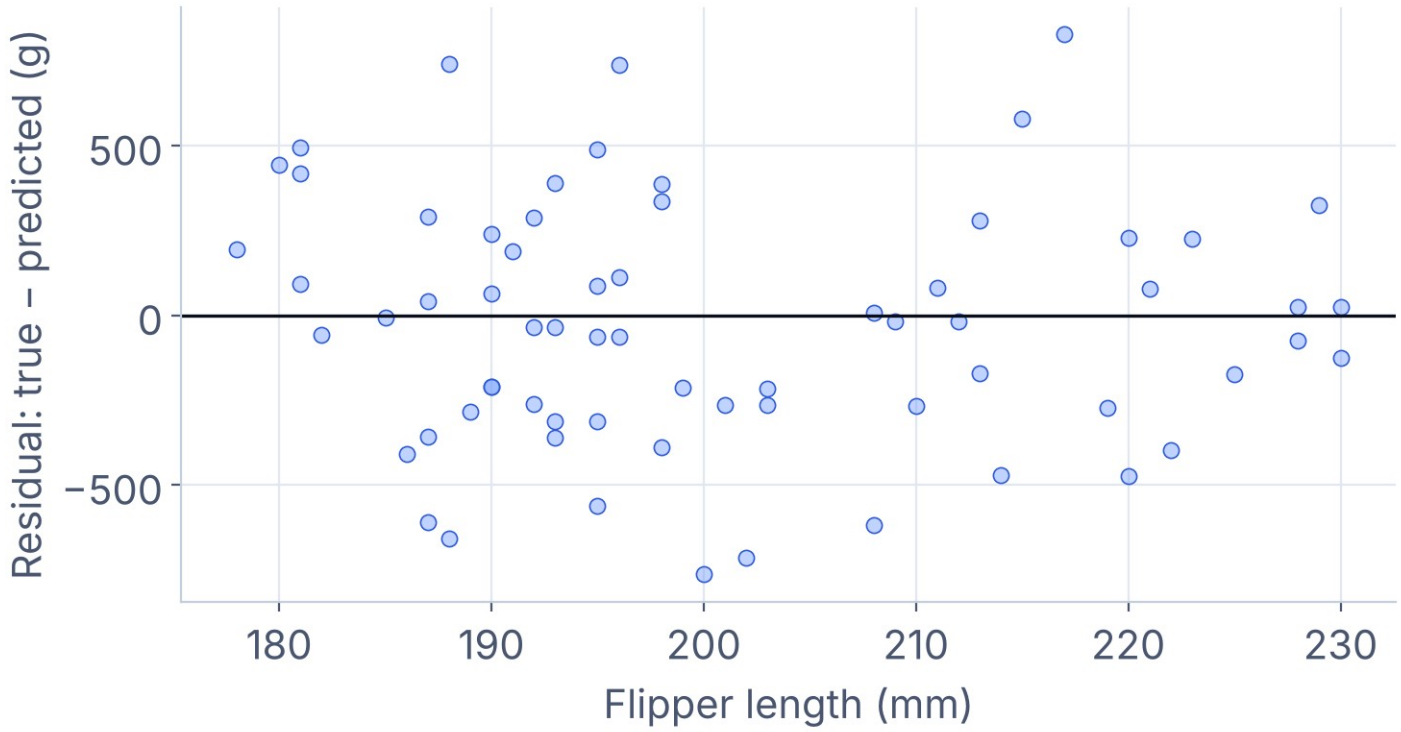

Figure 1. Test residuals of the flipper-only model (67 penguins). scikit-learn 1.9.1.

WHAT TO NOTICE

- Points on both sides of 0
- No curve, no funnel
- Largest errors about ±800 g

B

PART

Adding categories: species and sex

The same feature, a different slope: why the other features matter.

PART B · OBSERVATION

The bill depth paradox

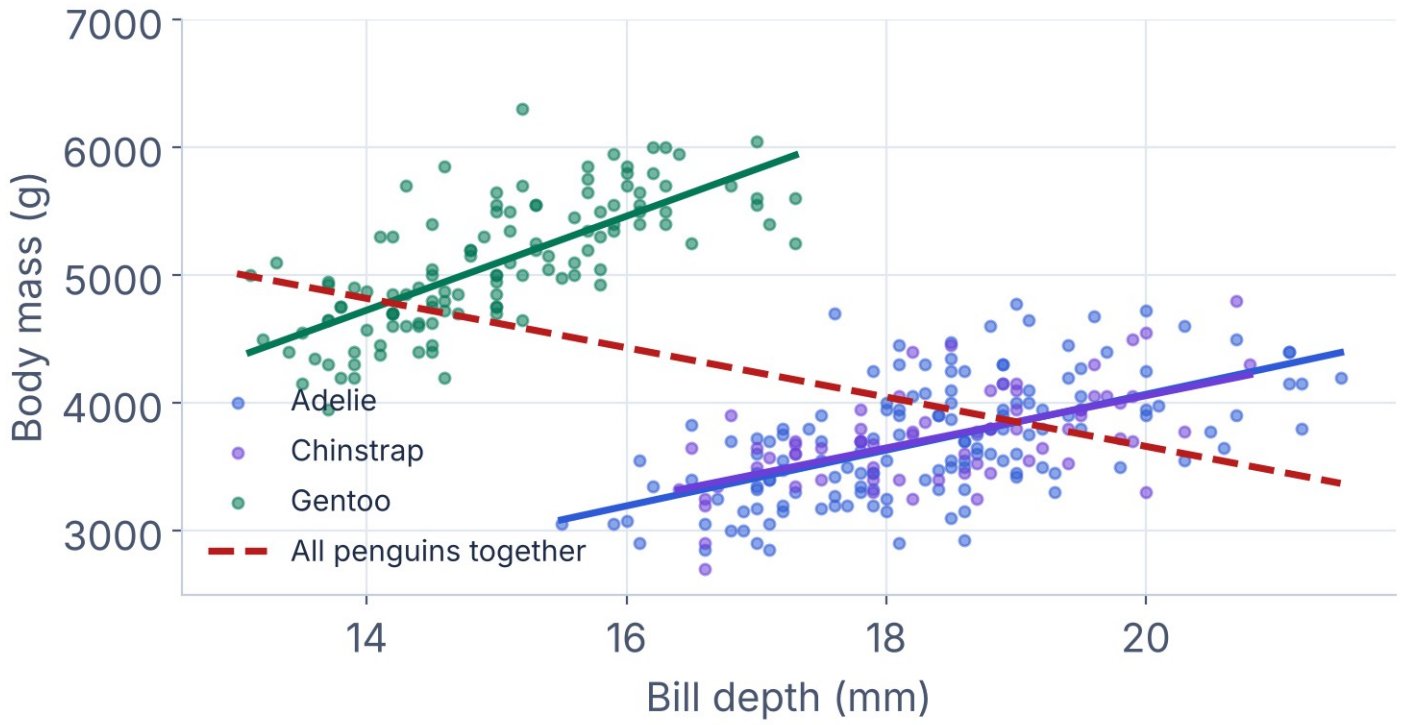

Figure 2. Penguins (n = 333). Lines: least squares fitted per species and on all penguins.

WHAT TO NOTICE

- All together: the slope is negative
- Within each species: positive
- Gentoo: heavy, shallow bills

PART B · GUIDED CODE

Part B · A pipeline with categories

CODE · PYTHON

```python
from sklearn.pipeline import make_pipeline
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import OneHotEncoder

prep = ColumnTransformer([
  ("num", "passthrough", ["bill_depth_mm"]),
  ("cat", OneHotEncoder(drop="first"), ["species"])])
model = make_pipeline(prep, LinearRegression())
cols = ["bill_depth_mm", "species"]
model.fit(train[cols], y_tr)
print(model[-1].coef_)
```

OUTPUT

```
[ 261.9
    2.9
 2277.1]
```

Bill depth, Chinstrap, Gentoo (reference: Adelie). Rounded.

PART B · EXERCISE

Part B · Questions

GOAL

See how the coefficient of a feature depends on the other features.

TIME

15 min

QUESTIONS

1. Fit bill depth alone, then with species: slopes and test R^2
2. Interpret the bill depth and Gentoo coefficients
3. Add flipper, bill length and sex: MAE, RMSE, R^2; read the Male coefficient
4. Predict a Gentoo female: 210 mm flipper, bill 45.0 × 15.0 mm

DELIVERABLE

One sentence per question, "all else being equal".

PART B · CHECK

Parts A and B · Check your results

Model	MAE (g)	RMSE (g)	R^2 test	R^2 train
Baseline (training mean)	674.1	794.6	−0.002	0.000
Flipper length	290.1	360.4	0.794	0.754
Bill depth	593.9	726.6	0.162	0.237
Bill depth + species	275.1	349.3	0.806	0.801
3 measurements + species + sex	196.5	253.9	0.898	0.869

Table 1. Penguins: 266 training, 67 test (random_state=42). scikit-learn 1.9.1.

C

PART

Restaurant tips, on your own

How much will a customer tip? Same workflow, new data.

PART C · DATA

The tips dataset

Column	Type	Meaning	Example
tip	number (target)	Tip left, in dollars	1.01
total_bill	number	Bill amount, in dollars	16.99
size	number	People at the table	2
sex	category	Sex of the person who paid	Female
smoker	category	Smokers at the table	No
day / time	category	Thur to Sun / Lunch or Dinner	Sun / Dinner

Table 2. tips.csv, 244 rows. Source: Bryant & Smith (1995), via the seaborn-data repository.

PART C · EXERCISE

Part C · Predict the tip

GOAL

Predict tip from tips.csv and choose the best model on the test set.

TIME

30 min

QUESTIONS

1. Read tips.csv: rows, mean tip, correlation bill-tip
2. Split (0.2, random_state=42) and baseline scores
3. Model with total_bill: interpret slope and intercept
4. Add size, then all columns (one-hot the categories)
5. Table of MAE, RMSE, R^2 for every model: which one wins?

DELIVERABLE

Your comparison table and two sentences of conclusion.

PART C · CHECK

Part C · Check your results

Model	MAE ($)	RMSE ($)	R^2 test	R^2 train
Baseline (training mean 3.09 $)	1.047	1.204	−0.159	0.000
total_bill	0.621	0.754	0.545	0.431
total_bill + size	0.664	0.805	0.481	0.452
All columns (one-hot)	0.667	0.839	0.437	0.458

Table 3. Tips: 195 training, 49 test (random_state=42). scikit-learn 1.9.1.

PART C · CORRECTION

Believing that more features always help

× INCORRECT APPROACH

"The model with all columns has the best training R^2 (0.458), so it is the best model"

✓ CORRECTION

Choose on the test set: total_bill alone has the best test R^2 (0.545) and the lowest MAE (0.62 $)

WHY IT FAILS

Adding features always raises the training R^2. On new bills, the extra columns add noise rather than information.

PART D · BONUS

Part D · Bonus, if you finish early

GOAL

Look at tips as a percentage, and test the limits of the model.

TIME

10 min

QUESTIONS

1. Create tip_pct = 100 × tip / total_bill: mean, median, max
2. Compare the slope of C3 with the median percentage
3. Predict the tip for a 100 $ bill. Can you trust it?
4. Plot the residuals of C3 against total_bill: what shape?

DELIVERABLE

Answers in Markdown cells under each plot.

Before you leave: checklist

CHECKLIST

- Notebook runs from top to bottom (Restart and Run All)
- Baseline computed for each dataset
- Every slope interpreted with units
- Comparison table for tips, with a conclusion
- Saved as Practice02_Lastname.ipynb and exported to HTML

NEXT

- Finish Part C at home if needed
- Next practice: classification with logistic regression

Resources

(1) James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). An Introduction to Statistical Learning, with Applications in Python, ch. 3 (Linear regression). Springer. statlearning.com

(2) scikit-learn 1.9 User Guide: Linear models; ColumnTransformer; Regression metrics. scikit-learn.org/stable

(3) Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. doi.org/10.5281/zenodo.3960218

(4) Bryant, P. G., & Smith, M. A. (1995). Practical Data Analysis: Case Studies in Business Statistics. Irwin. Tips data via github.com/mwaskom/seaborn-data

Course page
learn.drsoufiane.com

SH

Good work · Questions?

Keep your notebook: the next practice reuses the same workflow for classification.

Dr. Soufiane Hourri, PhD, HDR

drsoufiane.com