Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

CHOOSING A REGRESSION

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl

COMPLEXITY OF REGRESSION

- Linear regression is simpler than polynomial regression.
- Polynomial regression of degree 2 is simpler than degree 3, etc…
  $y = a_1\mathbf{x} + a_2\mathbf{x}^2 + b$ is simple than $y = a_1\mathbf{x} + a_2\mathbf{x}^2 + a_3\mathbf{x}^3 + b$
- Complex regressions have more risks of over-fitting.
- Simpler regressions have more risk of under-fitting.
- Choose the complexity by trying simpler to complex regressions.
  Choose the lowest complexity: the one after which there negligible improvements of the residuals (their magnitude & distribution).

CATEGORICAL DATA AS PREDICTOR

- **2 categories**: do nothing (encode them with 0 and 1).
- **Unordered categories**: Dummy code.
- **Ordered categories** (ordinal data):
  - Few classes (e.g., 3), large bins, or heteroscedasticity: Dummy code.
  - Many classes, small bins, homoscedasticity: do nothing.

PREDICTING CATEGORICAL DATA

- **2 categories**: No dummy code (encoding with 0 and 1).
  Use **logistic regression**.

- **Unordered categories**: Use **logistic regression**.
  Use a regression model for each class.
  Select the class with highest probability.

- **Ordered categories** (ordinal data):
  - Few classes (e.g., 3), large bins, or heteroscedasticity: same as unordered categories.
  - Many classes, small bins, homoscedasticity: try with linear/polynomial regression

Amsterdam University of Applied Sciences | DIGITAL SOCIETY SCHOOL

QUESTIONS?

EMMA BEAUXIS-AUSSALET

e.m.a.l.beauxis@hva.nl