“The world (of learning) is not linear!”
Statistical methods beyond linearity

Valérie Chavez-Demoulin

Faculty of Business and Economics,
University of Lausanne, Switzerland

October 9, 2014
LAW-2014, EPFL
Afternoon’s Agenda

I. Explaining one variable using several covariates: From linear models to most flexible models
   I.1. Linear models (cf. Łukasz’s course)
   I.2. Generalized linear models
      ▶ Practical in R on MOOC dataset
   I.3. Generalized additive models
      ▶ Practical in R on MOOC dataset

II. Explaining the dependence structure between two variables: From linear dependence structure, to non-linear dependence patterns
   II.1. Pearson correlation and its limitations
   II.2. Copula-based models
      ▶ Practical in R on MOOC dataset
PART I: One variable, several covariates
Linear models: A review

In a general\(^1\) linear model

\[ y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i \]

the response \( y_i, i = 1, \ldots, n \), is modelled by a linear function of the covariates \( x_j, j = 1, \ldots, p \), plus an error term \( \epsilon_i \).

Typically \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \) as a basis of inference e.g., \( t \)-tests on the parameters.

---

\(^1\)general here refers to dependence on potentially more than one covariate
Linear models: Four examples and a counter-example

A simple linear model: \( y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i \) (left panel)
A polynomial model: \( y_i = \beta_0 + \beta_1 x_{1i}^2 + \epsilon_i \) (right panel)

The models are linear in the parameters!

Other examples:
\[ y_i = \beta_0 + \beta_1 x_1 + \beta_2 \log(x_2) + \epsilon_i, \quad y_i = \beta_0 + \beta_1 x_3^{x_4} + \beta_2 x_5 + \epsilon_i \]
A Counter-example: \( y_i = \beta_0 + \beta_1 x_1^{\beta_2} + \epsilon_i \)
Linear models: in R

- Use the function `lm` for linear models (no specific package required) or function `gam` of the `mgcv` package
- Syntax for the formula:

  \[ y \sim x_1 + x_2 \]

  ▶ \( y \) is the name of the response vector
  ▶ \( x_1, x_2 \) are the names of the covariates (numerical vectors or factors)

- Other symbols can be used in the formula:
  ▶ \( x_1 : x_2 \) for an interaction component between \( x_1 \) and \( x_2 \)
  ▶ \( x_1 \times x_2 \) for compact version of \( x_1 + x_2 + x_1 : x_2 \)
  ▶ \( -1 \) to exclude an intercept
Example in R: MOOC dataset

It seems that there is an interaction between "groups of students" (strong or weak) and "week"!
R inputs and outputs

```r
> mod.lm <- gam(mvs ~ group*week-1, data=db)
> summary(mod.lm)

Family: gaussian
Link function: identity

Formula:
mvs ~ group * week - 1

Parametric coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| groupstrong | 1.056185 | 0.014898 | 70.895 | < 2e-16 *** |
| groupweak | 1.045786 | 0.014898 | 70.197 | < 2e-16 *** |
| week | 0.038268 | 0.002401 | 15.938 | 3.07e-11 *** |
| groupweak:week | -0.017469 | 0.003396 | -5.145 | 9.79e-05 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) = 0.959  Deviance explained = 100%
GCV = 0.0005945  Scale est. = 0.0004756  n = 20
```

The estimated model is

\[
\begin{align*}
\hat{E}(mvs_{i}^{\text{strong}}) &= 1.056 + 0.038 \text{week}_i, \\
\hat{E}(mvs_{i}^{\text{weak}}) &= 1.046 + \text{week}_i(0.038 - 0.017) = 1.045 + 0.021 \text{week}_i,
\end{align*}
\]
R outputs explained

- \( \Pr(>|t|) \): p-values (cf. Łukasz’s course)
- R-sq.(adj): \( R^2 \) adjusted for small size
- Deviance explained: serves as a generalization of R-squared
  \((\text{DevNull}-\text{DevResidual})/\text{DevNull}\)
- GCV (generalized Cross Validation): estimates the mean square prediction error
R graphical output

Use the function `predict(mod.lm,newdata,se.fit=T)` to get fitted values of the linear predictor for a new data set with corresponding standard error estimates.

![Graph showing mean video average speed over weeks for strong and weak students.](image)
Generalized linear models: A review

The **generalized** linear model (GLM) framework consists of four elements:

- A linear predictor \( \eta_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} \)
- A link function \( h \) that describes how the mean \( E(Y_i) = \mu_i \) depends on the covariates
  \[ h(\mu_i) = \eta_i \]
- A variance function \( V \) that describes how the variance, \( \text{var}(Y_i) \)
  depends on the mean
  \[ \text{var}(Y_i) = \phi V(\mu_i) \]
- A distribution from the *exponential family*

\[ ^2 \text{generalized here refers to an extension of the general linear model framework to variables that are not normally distributed} \]
### Generalized linear models: Three important families

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Response</th>
<th>Expectation</th>
<th>Link name</th>
<th>Link function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal⁴</td>
<td>$Y_i \sim \mathcal{N}(\mu_i, \sigma_i)$</td>
<td>$E(Y_i) = \mu_i$</td>
<td>Identity</td>
<td>$\eta_i = \mu_i$</td>
</tr>
<tr>
<td>Poisson</td>
<td>$Y_i \sim \mathcal{P}(\lambda_i)$</td>
<td>$E(Y_i) = \lambda_i$</td>
<td>Log</td>
<td>$\eta_i = \ln(\mu_i)$</td>
</tr>
<tr>
<td>Binomial</td>
<td>$Y_i \sim \mathcal{B}(n_i, p_i)$</td>
<td>$E(Y_i/n_i) = p_i$</td>
<td>Logit</td>
<td>$\eta_i = \ln\left(\frac{p_i}{1-p_i}\right)$</td>
</tr>
</tbody>
</table>

³ this is the general linear model
Generalized linear models in R

- Use the `glm` function (no specific package required) or `gam` function of the `mgcv` package.
- Specify the family function using the argument `family`:
  - `gaussian` is by default; it uses the identity link.
  - `poisson`; it uses the log link.
  - `binomial`; it uses the logit link.
  - ...
- `glm` returns an object of class “glm”. Assess your fit using:
  - `summary()` to assess significant components.
  - `residuals()` to check the residuals (any pattern?).
  - `predict()` to get predicted values from your GLM estimated model.
  - `coeff()`, `fitted()`, `deviance()`, ...
Example in R: MOOC dataset

Are Belgian students more active than Swiss students? (We consider only students who got a “distinction” to remove the interaction effect between student level and country)
Transform the variable $\log(X + 1)$ to better show the difference

Are Belgian students more active than Swiss students?
Let $R$ denote the “Total forum view”. It is a counting variable so we suppose that $R \sim \mathcal{P}(\lambda)$.

Call:

\[
\text{glm(formula = totalForumView} \sim \text{Country + Course, family = poisson, data = db)}
\]

Deviance Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-11.795</td>
<td>-7.872</td>
<td>-4.852</td>
<td>-0.335</td>
<td>161.615</td>
</tr>
</tbody>
</table>

Coefficients:

|             | Estimate  | Std. Error | z value | Pr(>|z|) |
|-------------|-----------|------------|---------|----------|
| (Intercept) | 3.165679  | 0.040026   | 79.084  | < 2e-16 *** |
| CountryBelgium | 0.928940  | 0.043954   | 21.134  | < 2e-16 *** |
| CountryCanada | 0.939264  | 0.043253   | 21.715  | < 2e-16 *** |
| CountryFrance | 0.959978  | 0.040018   | 23.989  | < 2e-16 *** |
| CountryMorocco | 0.247794  | 0.047544   | 5.212   | 1.87e-07 *** |
| CountrySpain | 0.540451  | 0.043139   | 12.528  | < 2e-16 *** |
| CountrySwitzerland | 1.076676  | 0.040491   | 26.591  | < 2e-16 *** |
| CountryUnited States | 0.417509 | 0.043117   | 9.683   | < 2e-16 *** |
| Coursejava | -0.053286 | 0.008669   | -6.147  | 7.91e-10 *** |

The estimated model is $\log \hat{\lambda} = 3.16(0.040) + 0.93(0.044)I_{Belgium} + \cdots + 0.42(0.043)I_{US} + \cdots - 0.053(0.0087)I_{java}$

For instance $\hat{E}(R_{US,C++}) = \exp(3.16 + 0.42) \approx 36$
R output on explaining total post on thread

Let $R$ denotes the “Total post on thread” and suppose that $R \sim \mathcal{P}(\lambda)$

Call:
```
glm(formula = totalPostonThread ~ Country + Course, family = poisson,
data = db)
```

Deviance Residuals:
```
Min 1Q Median 3Q Max
-3.149 -2.171 -1.914 -0.367 49.525
```

Coefficients:
```
                Estimate Std. Error  z value Pr(>|z|)
(Intercept) 0.60555 0.16013   3.781 0.000156 ***
CountryBelgium -0.28236 0.20951  -1.348 0.177750
CountryCanada 0.41940 0.18071   2.321 0.020295 *
CountryFrance 0.52629 0.18089   3.271 0.001071 **
CountryMorocco 0.20836 0.19039  1.094 0.273796
CountrySpain 0.05018 0.18036   0.278 0.780827
CountrySwitzerland 0.99519 0.16191   6.147 7.92e-10 ***
CountryUnited States -0.11740 0.18149  -0.647 0.517723
Coursejava -0.36882 0.04003  -9.214 < 2e-16 ***
```

Example: $\hat{E}(R_{Belgium,java}) = \exp(0.60 - 0.28 - 0.37) \approx 0.95$

- Swiss students are the most active
- Belgium, Morocco, Spain and US can be grouped together
- Students in JAVA course are less active than students in C++ course
Practical in R

1. Load the data `tsstudent`
2. Remove students who dropped out (grades = 0)
3. Plot the histogram of the grades and comment
4. Add an indicator variable for “Passed” (grade ≥ 75) or “Failed”
5. Draw some plots
6. Provide a relevant model for the probability of passing as a function of significant covariates
7. Randomly choose 99% of the students in the sample (non-dropouts) and fit the model of point 6
8. Predict the “Passed” or “Failed” on the remaining 1% of students left (based on the predicted probability estimated from the model)
9. Repeat points 7 and 8 \( B = 100 \) times and obtain a confidence interval for the prediction

The Monte Carlo procedure in point 9 is called Bootstrap (cf Łukasz’s course)
More flexibility is sometimes needed

The recorded head acceleration (in g)

Time since impact (in millisenonds)
Generalized additive models

Extend GLM to non-parametric form:

\[ h(\mu) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p) \]

where \( f_j, j = 1, \ldots, p \) are non-parametric functions

Semi-parametric models are also possible

\[ h(\mu) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{r-1} x_{r-1} + f_r(x_r) + \ldots + f_p(x_p) \]
Example in R: Pisa dataset

The data are average Science scores by country from the Programme for International Student Assessment. The key variables are:

- Overall Science Score (average score for 15 year old)
- Interest in science
- Support for scientific inquiry
- Income Index
- Health Index
- Education Index
- Human Development Index (composed of the Income Index, Health Index, and Education Index)

---

4 An R data package for the Programme of International Student Assessment (PISA)
Univariate and bivariate densities
Comment on univariate and bivariate densities

- The univariate distributions look “roughly” normally distributed (not always)
- So that, bivariate distributions look “roughly” multinormal (not always)
- Negative correlation between interest and overall score for science (Simpson’s paradox!)
- Some correlations are significant but the dependence structure is not always linear
Dependence of overall score on four covariates
Let’s try a linear model to explain overall score in terms of income

```r
> library(pisa)
> mod.LM <- gam(Overall ~ Income, data = pisa)
> summary(mod.LM)

Family: gaussian
Link function: identity

Formula:
Overall ~ Income

Parametric coefficients:

|            | Estimate | Std. Error | t value | Pr(>|t|) |
|------------|----------|------------|---------|----------|
| (Intercept)| 204.32   | 35.37      | 5.777   | 4.32e-07 *** |
| Income     | 355.85   | 46.79      | 7.606   | 5.36e-10 *** |

---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) = 0.518  Deviance explained = 52.7%
GCV = 1504.5  Scale est. = 1448.8  n = 54
```
Does the linear model look meaningful?
“Let the data speak for themselves”: fit a GAM

```r
> library(mgcv)
> mod.GAM <- gam(Overall ~ s(Income), data = pisa)
> summary(mod.GAM)

Family: gaussian
Link function: identity

Formula:
Overall ~ s(Income)

Parametric coefficients:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| (Intercept) | 470.444    | 3.968   | 118.6    | <2e-16 *** |

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

<table>
<thead>
<tr>
<th>edf</th>
<th>Ref.df</th>
<th>F</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s(Income)</td>
<td>7.972</td>
<td>8.729</td>
<td>15.65</td>
</tr>
</tbody>
</table>

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) = 0.717   Deviance explained = 75.9%
GCV = 1019.7  Scale est. = 850.26  n = 54
```
More on GAM

- GAM uses essentially *cubic spline*
- A cubic spline is a connection of multiple cubic polynomial regressions
- They are joined at some *knots* to create a continuous curve
- $s(\ldots, k)$ is the smooth function in which the type of smoothing is specified
- $k$ if specified, it determines the degrees of freedom (number of terms associated with the smoother)
- edf is the *effective degrees of freedom* (related to $k$ if specified)
- More on `?summary.gam`
Get the resulting smoothed curve

$\hat{E}(\text{overall} \mid \text{Income} = 0.6) = 470.4 - 70 \approx 400$
Check the residuals

gam.check(mod.GAM)
How to compare models?

Use AIC! (cf. Łukasz’s course)

- AIC for Pisa linear model

\[
\text{mod.LM\$aic} \\
550.2449
\]

- AIC for Pisa GAM

\[
\text{mod.GAM\$aic} \\
527.6372
\]

Choose the model with lower AIC
How to choose the degrees of freedom $k$?

Use AIC!
Practical in R

Evolution of students ("Passed" and "Failed") behaviour over time

![Graphs showing mean assignment submission, lecture view, post on thread, and forum view over time.](image-url)
1 Get the data PostonThreadWeek
2 Build several models for the expected number of “Post on Thread” as a function of the covariates “Success” (defining students who passed (1) or failed (0)) and “Week”
   - start with a linear model for “Week”
   - consider smoothed models for “Week” by increasing the degrees of freedom (k) from 2 to 9
3 Plot the AIC against the degrees of freedom, comment and select a model
4 Plot the fitted expected number of “Post on Thread” against “Week” for the two groups of students (passed or failed)
PART II: Two variables, a dependence structure
Understanding limitations of Pearson correlation

- Not appropriate for non linear phenomena
- Some technical limits:
  - not robust to outliers or extremes
  - variances of the two variables must exist
  - not invariant to some transformations

\[
\text{correlation}=0.70
\]

\[
\text{correlation}=0.55
\]
Log-transformation can make the dependence more linear.

**Model on log-scale:** \( \ln(y) = 3.33(0.14) + 0.20(0.04) \ln(x) \)

**Interpretation on original scale:** \( y = 27.9x^{0.20} \)
 Conditioning can sometimes highlight dependence structures
In MOOC, dependence phenomena are often non linear!
How to handle complicated dependence structures?

Our aim is to model $F(x_1, x_2) = P(X_1 \leq x_1, X_2 \leq x_2)$ when the dependence structure between $X_1$ and $X_2$ is non-linear.
Beyond correlation: Copula-based models

- Copula-based models are used to build the joint distribution of several \((d)\) variables.
- For the purpose of this course we consider \(d = 2\) so that we aim at modelling \(F(X_1, X_2)\), that is

\[
P(X_1 \leq x_1, X_2 \leq x_2)
\]

- Usual models consider bivariate distributions which impose same marginal distributions \(F_1\) and \(F_2\). The multinormal distribution, for instance, supposes that the two marginal distributions \(F_1\) and \(F_2\) are normal! This is sometimes restrictive.
- Most current bivariate distributions (normal, Student \(t\)) are based on correlation as measure of dependence. This is also restrictive (cf. previous slides)!
Copulas allow to model flexible joint distributions

One may want a model for $F(X_1, X_2) = P(X_1 \leq x_1, X_2 \leq x_2)$ which allows:

1. arbitrary marginal distributions. Example: $F_1$ is normal and $F_2$ Student
2. complicated dependence structure fitted separately from the marginal distributions
3. to ‘couple” together the marginals of point 1 and the dependence structure of point 2 to form a joint distribution $F(X_1, X_2) = P(X_1 \leq x_1, X_2 \leq x_2)$

→ Copula based models allow to build joint distribution with different marginal distributions and to handle complicated dependence structure !!
On a more theoretical aspect of copulas

An important theorem (Sklar, 1959) says that if the marginal distributions $F_1, F_2$ are continuous then there exists a unique copula $C$ such that for all $x_1, x_2$ in $[-\infty, \infty]$

$$F(x_1, x_2) = C\{F_1(x_1), \ldots, F_d(x_d)\}$$

And conversely, if $C$ is a copula and $F_1, F_2$ are univariate distribution functions, then $F$ defined above is a multivariate distribution function with margins $F_1, F_2$
Copulas and dependence structures

Sklar’s theorem shows

- that any bivariate distribution contains a copula
- how a bivariate distribution can be formed by coupling together a copula and the marginal distributions

We sometimes refer to \( C \) as the dependence structure of \( F \)

Invariance

\( C \) is invariant under strictly increasing transformations of the marginals.

If \( T_1, \ldots, T_d \) are strictly increasing, then \((T_1(X_1), \ldots, T_d(X_d))\) has the same copula as \((X_1, \ldots, X_d)\)
Copula model highlights a fallacy in the use of correlation

Consider the random vector \((X_1, X_2)\)

“Marginal distributions and correlation determine the joint distribution”

- True for the class of bivariate normal distributions or, more generally, for elliptical distributions
- Wrong in general, as the next example shows
Gaussian and Gumbel copulas compared

Gaussian Copula, rho = 0.7

Normal Q–Q Plot

Normal Q–Q Plot

Gumbel Copula, rho = 0.7

Normal Q–Q Plot

Normal Q–Q Plot

Margins are standard normal; correlation is 70%
Examples of copulas

Clayton copula

Gumbel copula
Practical in R

How do we model the dependence structure between these two variables of student activity?

(the red line shows how poor is a linear model in that case)
Practical in R

1. Get the data tsstudent
2. Consider the two variables: $Y_1$ the aggregated number of forum view over week (for students with grade=100) and $Y_2$ the aggregated number of post on thread over week (for students with grade=100)
3. Make a scatterplot of the sample observations $(y_{i1}, y_{i2}), i = 1, \ldots, n$, where $n$ is the sample size
4. Consider $X_j = \log(Y_j + 1), j = 1, 2$ and scatterplot $(x_{i1}, x_{i2}), i = 1, \ldots, n$. Comment the dependence pattern and discuss the accuracy of a multinormal distribution for $(X_1, X_2)$.
5. Recall that the copula $C$ is the distribution of $(F_1(X_1), F_1(X_2))$. Estimate the pseudo-sample $\{u_{i1} = \hat{F}_1(x_{i1}), u_{i2} = \hat{F}_2(x_{i2})\}, i = 1, \ldots, n$ using the empirical distributions $\hat{F}_j, j = 1, 2$
6. Install and load copula package
7. Fit different copulas on the pseudo sample (Gaussian, Student $t$, Clayton, Gumbel)
8. Compare the copula models using AIC
9. Use your selected copula model to estimate $P(X_1 \leq x_1, X_2 \leq x_3)$, for some $(x_1, x_2)$ with $x_1 \geq 0, x_2 \geq 0$
First, look at the data! (it is easier to guess an indicator like the correlation from a scatterplot of the data than to guess the data pattern from an indicator)

Check the distribution of the response variable (normal/symmetric enough)?

If the dependence on covariate is not linear use a flexible nonparametric form: “let the data speak for themselves”

Fit and compare several models and select the best

Check your model and be objective (“no model is perfect”)

The statement “it is not possible to get any information from these data”, is in (and of) itself important information!
The take home message: Part II

- First point of Part I remains crucial: look at the marginal distributions of your variables $X_1$ and $X_2$ and look at their dependence structure.
- Estimate the univariate marginal distributions $F_1$ and $F_2$ of your variables $X_1$ and $X_2$ separately (using parametric models or empirically).
- Fit different copulas on the probability transformed variables $\hat{F}_1(x_1)$ and $\hat{F}_2(x_2)$, compare them and select the best $\hat{C}$.
- Estimate the joint probability of $X_1$ and $X_2$ using Sklar’s Theorem:

$$\hat{F}(x_1, x_2) = \hat{C}\{\hat{F}_1(x_1), \hat{F}_2(x_2)\}$$

- The copula-based model can be used for more than 2 variables!
References