PSY9511: Seminar 5

Beyond linearity: Extensions of linear models and tree-based models

Esten H. Leonardsen

27.10.24

UNIVERSITETET
I OSLO

Outline

Outline

1. Exercise 3
2. Exercise 4
3. Recap
4. Extensions of linear models
   - 4.1 Generalized linear models (GLMs)
   - 4.2 Generalized additive models (GAMs)
5. Tree-based models
   - 5.1 Decision trees
   - 5.2 Random forests
   - 5.3 Gradient boosting (XGBoost)
6. Exercise 5

Exercise solutions

https://uio.instructure.com/courses/59550

Exercise 3: Solution

http://localhost:8888/notebooks/notebooks/Solution%203.ipynb

Exercise 4

Exercise 4: Stratification

http://localhost:8889/notebooks/notebooks%2FStratification.ipynb

Exercise 4: Solution

http://localhost:8888/notebooks/notebooks/Solution%204.ipynb

Recap

UNIVERSITETET
I OSLO

Recap

Whenever a modelling choice is made on the basis of performance in a dataset, **we have to assume the performance achieved by the chosen model is inflated**

! Don't use a single train/validation split, cross-validation or bootstrapping with only train/validation sets to select the best model and then reports its performance

→ Hold out a test set

→ Use a nested cross-validation

Extensions of linear models

Non-linear models: Nothing to worry about

```r
formula <- ...
result <- glm(formula, family=Gamma(link="log"), data=data)
```

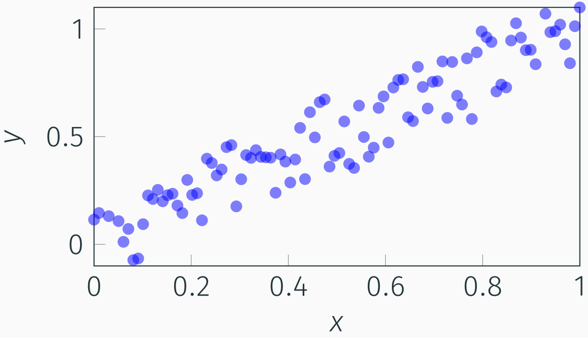
```r
formula <- ...
result <- gam(formula, data=data)
```

In[2]:
```python
from xgboost import XGBClassifier

model = XGBClassifier()
model.fit(X, y)
```

Extensions of linear models: Generalized linear models

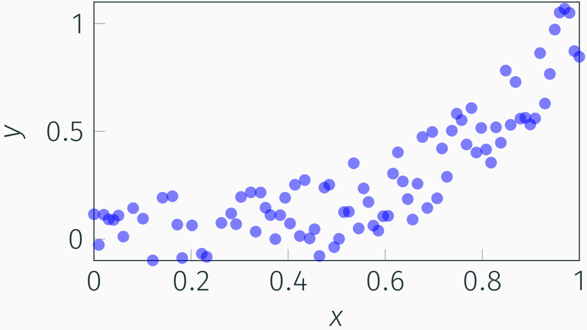
$$\hat{y} = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



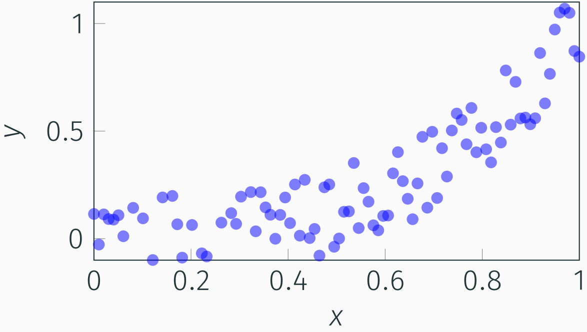
$$\hat{y} = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



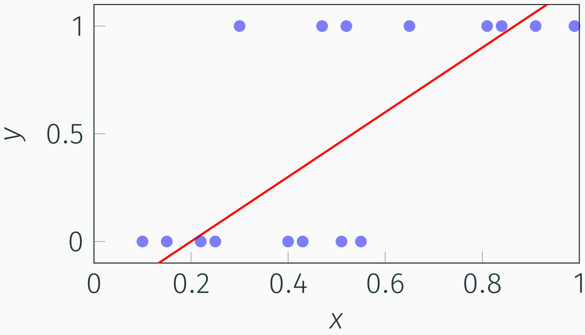
$$\hat{y} = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



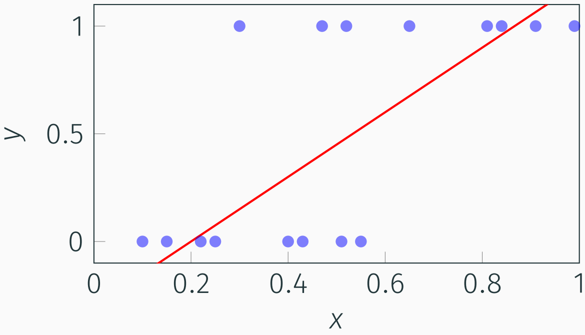
$$\hat{y} = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



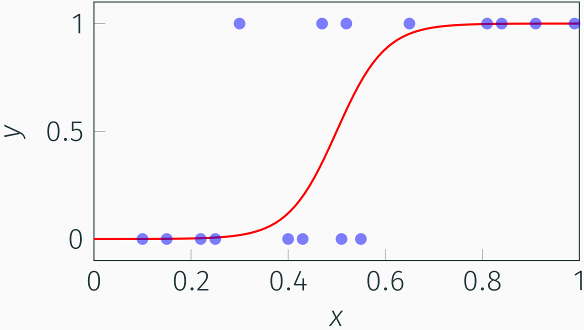
$$\hat{y} = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



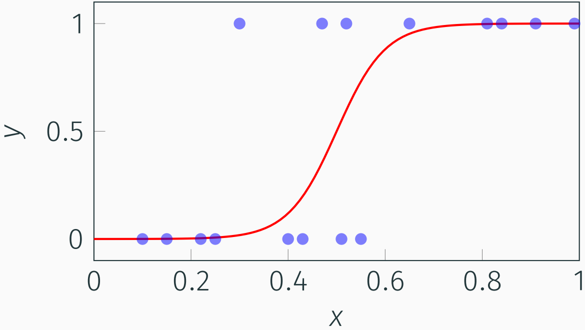
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



$$p(X) = \frac{e^{\left(\beta_0 + \sum_{i=0}^{p} \beta_i x_i\right)}}{1 + e^{\left(\beta_0 + \sum_{i=0}^{p} \beta_i x_i\right)}}$$

Extensions of linear models: Generalized linear models



$$p(X) = \frac{e^{\left(\beta_0 + \sum_{i=0}^{p} \beta_i x_i\right)}}{1 + e^{\left(\beta_0 + \sum_{i=0}^{p} \beta_i x_i\right)}}$$

Extensions of linear models: Generalized linear models

$$\hat{y} = f(\beta_0 + \sum_{i=0}^{p} \beta_i x_i)$$

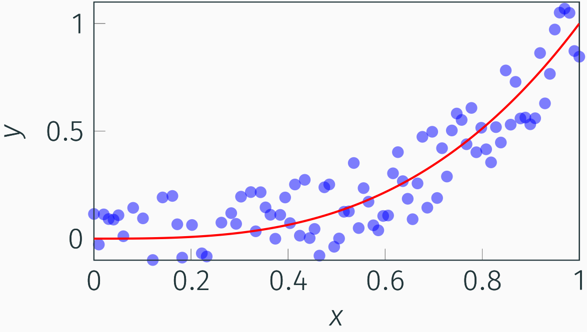
Extensions of linear models: Generalized linear models

$$f(\hat{y}) = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models

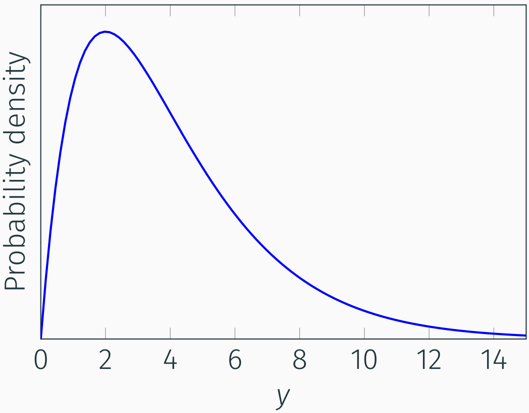
$$f(\hat{y}) = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



$$\log(\hat{y}) = \beta_0 + \sum_{i=0}^{p} \beta_i x_i$$

Extensions of linear models: Generalized linear models



Generalized linear models (GLMs):
Extends upon the regular linear model by associating the predictors to the response via a non-linear link function f.

- Requires us to specify f (often determined by investigating the distribution of the response).

Extensions of linear models: Generalized linear models

In[1]:

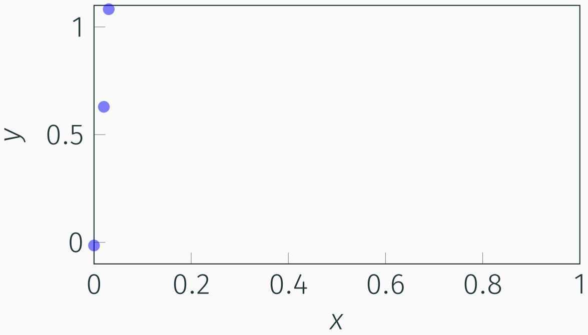
```
from sklearn.linear_model import GammaRegressor

model = GammaRegressor()
model.fit(X, y)
```

In[2]:

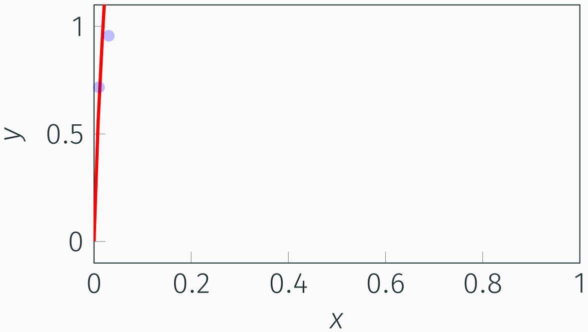
```
import statsmodels.api as sm

link = sm.genmod.families.links.Log()
model = sm.GLM(y, X, family=sm.families.Gamma(link=link))
model.fit()
```

```
formula <- ...
result <- glm(formula, family=Gamma(link="log"), data=data)
```

Extensions of linear models: Generalized additive models



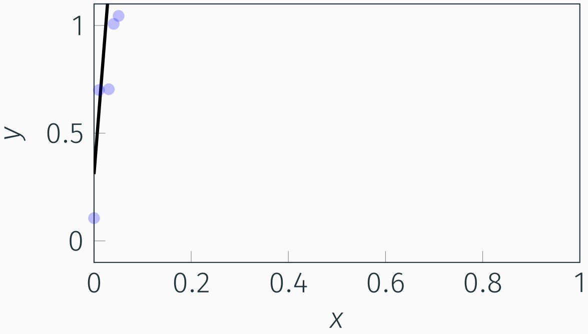
Extensions of linear models: Generalized additive models



Piecewise polynomial functions:

- Regression splines (ISL, Chapter 7.4)

Extensions of linear models: Generalized additive models



A single, complex, polynomial function:

- Smoothing splines (This lecture)

Extensions of linear models: Generalized additive models

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

Extensions of linear models: Generalized additive models

$$\hat{y}_i = g(x_i)$$

$$\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

Extensions of linear models: Generalized additive models

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

$$\uparrow$$

$$\sum(y_i - \hat{y}_i)^2$$

Extensions of linear models: Generalized additive models

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

$$\uparrow$$

Large when g is wiggly

Extensions of linear models: Generalized additive models

Balances what is most important

$$\underbrace{\sum_{i=1}^{n}(y_i - g(x_i))^2}_{\text{Tries to fit data}} + \overbrace{\lambda}^{\text{Balances what is most important}} \underbrace{\int g''(t)^2 dt}_{\text{Tries to simplify } g}$$

Tries to fit data

Tries to simplify g

Extensions of linear models: Generalized additive models

http://localhost:8888/notebooks/notebooks/Smoothing%20spline.ipynb

Extensions of linear models: Generalized additive models

$$\hat{y} = \beta_0 + \beta_1 x \qquad\qquad \hat{y} = g(x)$$

Extensions of linear models: Generalized additive models

$$\hat{y} = \beta_0 + \beta_1 x \quad\longrightarrow\quad \hat{y} = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

$$\hat{y} = g(x) \quad\longrightarrow\quad \hat{y} = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$$

Extensions of linear models: Generalized additive models

Generalized additive models (GAMs):
Extends upon the regular linear model by allowing for non-linear functions f_j to be fitted for each predictor x_j.

- Does not allow for interactions between predictors.

Extensions of linear models: Generalized additive models

scripts/gam.R