Imperial College London

Explainability

- The right to an explanation
- **Classical Interpretability and Partial Dependence Plots**
- An overview of XAI techniques
- Are all explanations causal?

Linearity and additivity are the basis of classic XAI

In [1]:
```
import pandas as pd
import shap
import sklearn
```

In [10]:
```
# the classic diabetes dataset from https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html
X,y = shap.datasets.diabetes()

# a simple linear model
lreg = sklearn.linear_model.LinearRegression()
lreg.fit(X, y)
```

Linearity and additivity are the basis of classic XAI

```
In [14]: print("Model coefficients:\n")
         for i in range(X.shape[1]):
             print(X.columns[i], "=", lreg.coef_[i].round(4))
```

```
Model coefficients:

age = -10.0122
sex = -239.8191
bmi = 519.8398
bp = 324.3904
s1 = -792.1842
s2 = 476.7458
s3 = 101.0446
s4 = 177.0642
```

Imperial College London

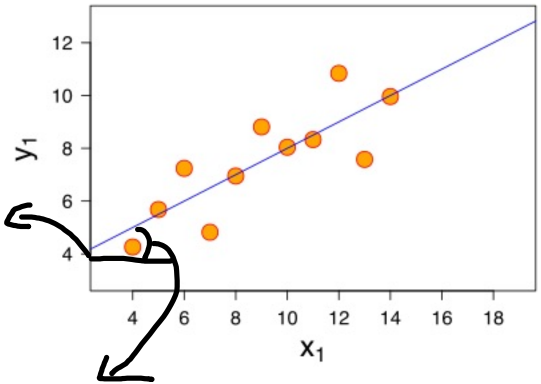
Linearity and additivity are the basis of classic XAI

$$y_i = \beta_0 + \sum_{j=1}^{n} \beta_j X_{ij}$$

```
In [40]: lreg.predict(np.array(X.loc[0]).reshape(1,-1))
Out[40]: array([206.11706979])

In [44]: np.sum(lreg.coef_*X.loc[0]) + lreg.intercept_
Out[44]: 206.11706978709228
```

intercept



Regression coefficient

https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Even linear regression requires careful interpretation

```
print("Model coefficients:\n")
for i in range(X.shape[1]):
    print(X.columns[i], "=", lreg.coef_[i].round(4))
```

```
Model coefficients:

age = -10.0122
sex = -239.8191
bmi = 519.8398
bp = 324.3904
s1 = -792.1842
s2 = 476.7458
s3 = 101.0446
s4 = 177.0642
s5 = 751.2793
s6 = 67.6254
```

```
lreg_no_s1 = sklearn.linear_model.LinearRegression()
X_no_s1 = X.drop('s1', 1)
lreg_no_s1.fit(X_no_s1, y)
print("Model coefficients (removing BMI):\n")
for i in range(X_no_s1.shape[1]):
    print(X_no_s1.columns[i], "=", lreg_no_s1.coef_[i].round(4))
```

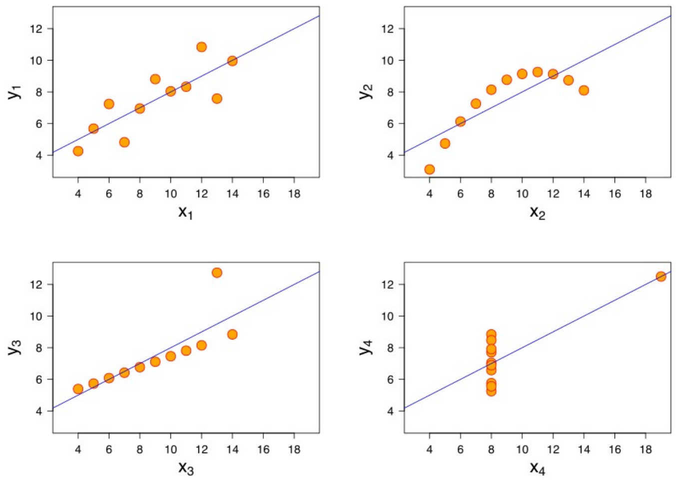
```
Model coefficients (removing BMI):

age = -7.9167
sex = -234.1587
bmi = 528.5262
bp = 319.7704
s2 = -143.2835
s3 = -250.5987
s4 = 70.4507
s5 = 461.8402
s6 = 69.126
```

Sign reversal can sometimes occur in linear regression when removing one variable from a multivariable regression

Even linear regression requires careful interpretation

- Explanations of the model's predictions do not necessarily explain the phenomenon itself.
- Checks of model assumptions are necessary in addition to checks of predictive accuracy for purposes of explainability / interpretability
- Robustness / stability is another key concern: communicated explanations should be qualitatively robust (e.g., top 5 most important drivers, signs, etc.) or, if not, supported by an explanation why not



Known as the Anscombe quartet, this selection of small datasets shows how the same line of best fit appears in 4 different situations

In tabular data, explainability involves four elements

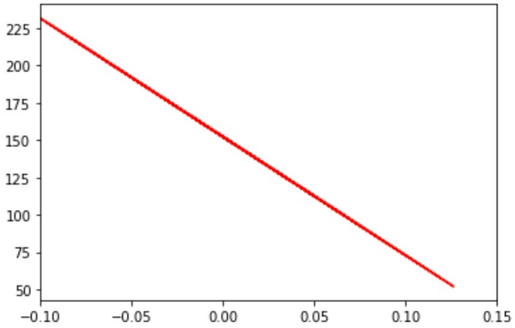
- Magnitude of feature importance
- Direction or shape of feature importance
- Interaction between features
- Confidence in feature importance

- Shap values
- Partial Dependence plots

Partial dependence plots

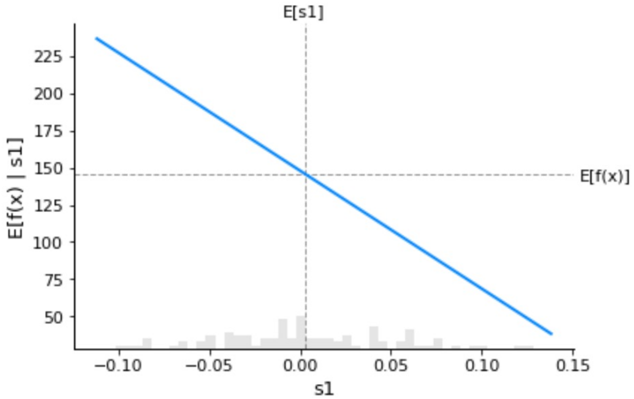
```
plt.plot(
    X100['s1'],
    lreg.coef_[4]*X100['s1'] + lreg.intercept_,
    color='red')
plt.xlim([-0.1,0.15])
```

```
(-0.1, 0.15)
```



$$p_2(\xi) = \hat{\beta}_2 \xi + \beta_0$$

```
X100 = shap.utils.sample(X, 100)
shap.plots.partial_dependence(
    "s1", lreg.predict, X100, ice=False,
    model_expected_value=True, feature_expected_value=True
)
```

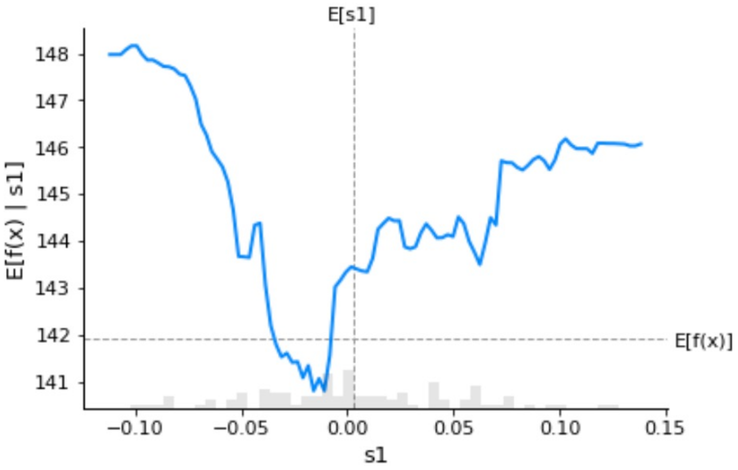


$$p_2(\xi) = \frac{1}{n} \sum_{i=1}^{n} f(x_{i1}, \xi, x_{i3}, \ldots)$$

Partial dependence plots

```
# a random forest
rf = sklearn.ensemble.RandomForestRegressor()
rf.fit(X, y)
shap.plots.partial_dependence(
    "s1", rf.predict, X100, ice=False,
    model_expected_value=True, feature_expected_value=True
)
```

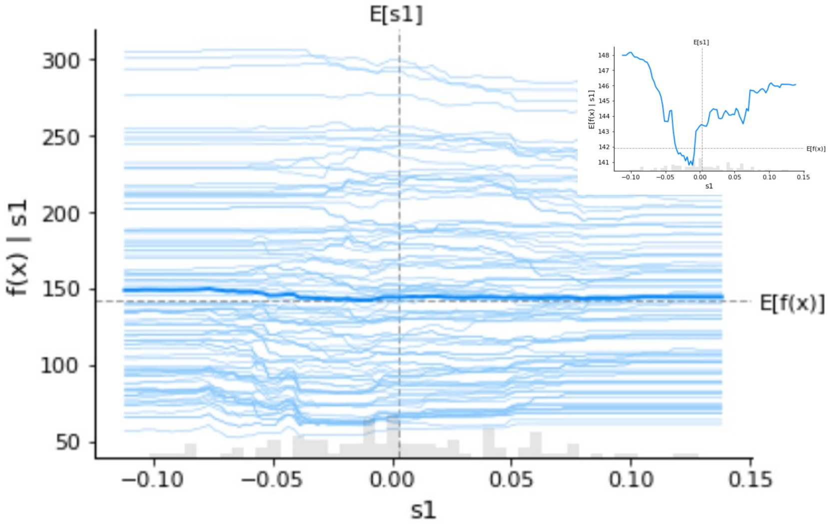
$$p_2(\xi) = \frac{1}{n}\sum_{i=1}^{n} f(x_{i1}, \xi, x_{i3}, \dots)$$



Individual Conditional Expectation (ICE) plots

$$p_2(\xi) = \frac{1}{n}\sum_{i=1}^{n} f(x_{i1}, \xi, x_{i3}, \dots)$$

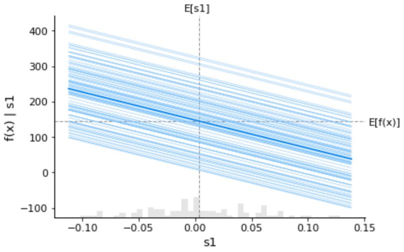
```
shap.plots.partial_dependence(
    "s1", rf.predict, X100, ice=True,
    model_expected_value=True, feature_expected_value=T
)
```



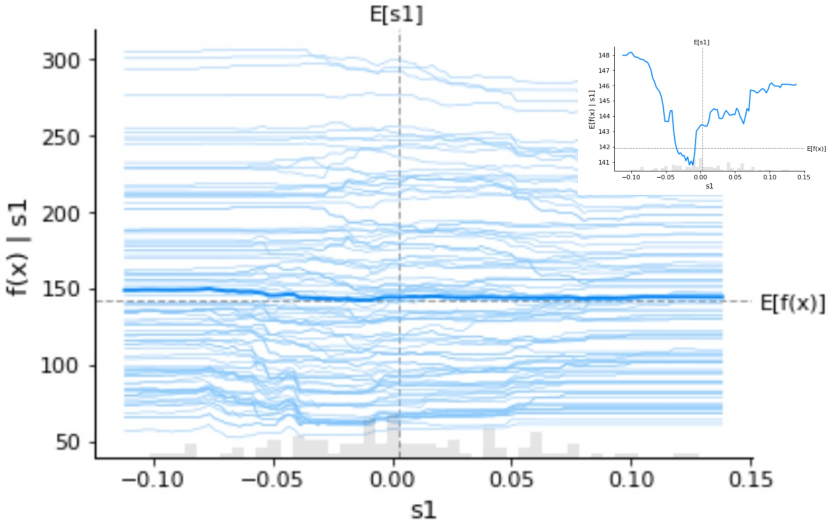
Individual Conditional Expectation (ICE) plots

$$p_2(\xi) = \frac{1}{n}\sum_{i=1}^{n} f(x_{i1}, \xi, x_{i3}, \dots)$$

For linear regression:




```
shap.plots.partial_dependence(
    "s1", rf.predict, X100, ice=True,
    model_expected_value=True, feature_expected_value=T
)
```



Imperial College London

Summary

- Classical interpretability relied on some mathematical properties like linearity and additivity.
- It produced explanations by way of assigning properties to individual features like “magnitude of effect”, “direction of effect”, ”interactions between features”, and “confidence”.
- Even simple methods like linear regression can however be misleading and require care.
- Modern XAI techniques attempt to offer similar “explanations” to classical methods by generalizing the concept of an “effect” through methods like partial dependence plots.
- Familiarity with these tools is an absolute must-have in the work of a practicing data scientist