REGRESSION

Introduction to linear regression models

Regression vs Classification

Predict a **continuous value** instead of discrete class

Regression

Weight	Color	Seeds	Price
150	80	8	2.5
200	112	6	3.1
170	120	8	2.9
210	105	7	3.6
180	130	9	2.4

attributes

Target is a continuous value

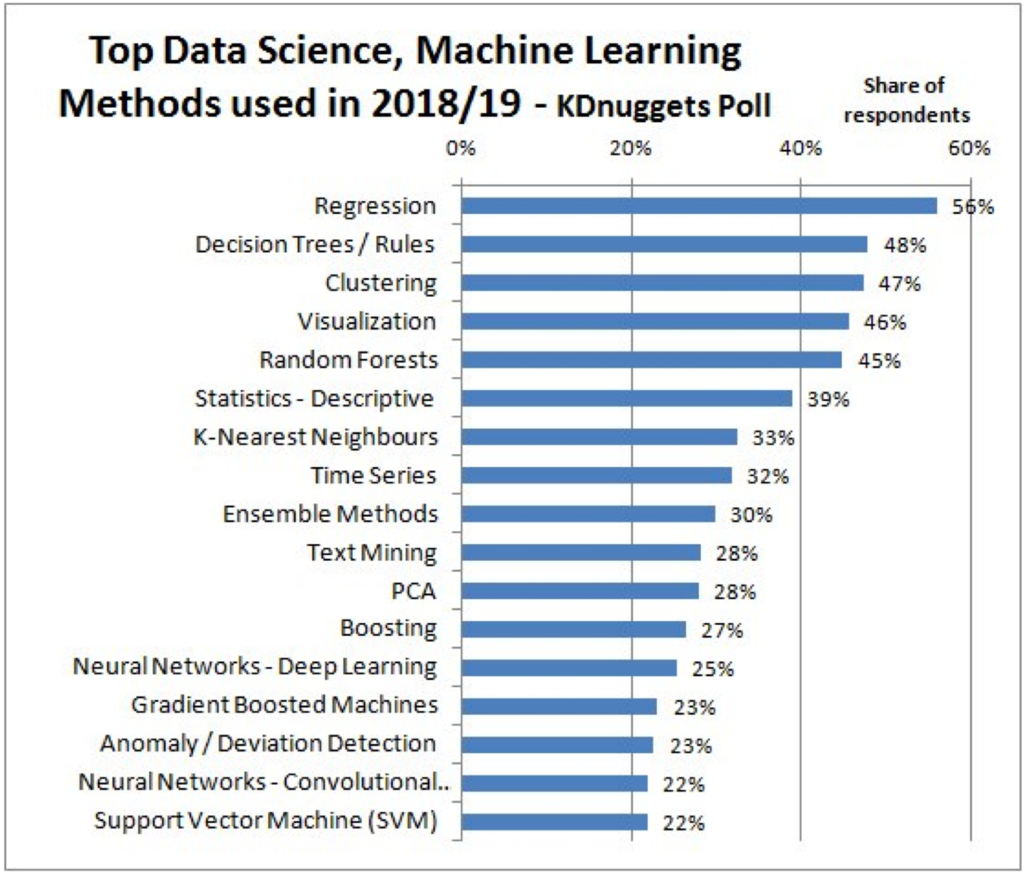
Classification

Weight	Color	Seeds	Fruit	
150	80	8	0	apple
200	112	6	1	orange
170	120	8	1	orange
210	105	7	1	orange
180	130	9	0	apple

attributes class (target)

- Both are **supervised** : model learned from a known training set
- Linear regression is the basis of a lot of models, and routinely used by data scientists

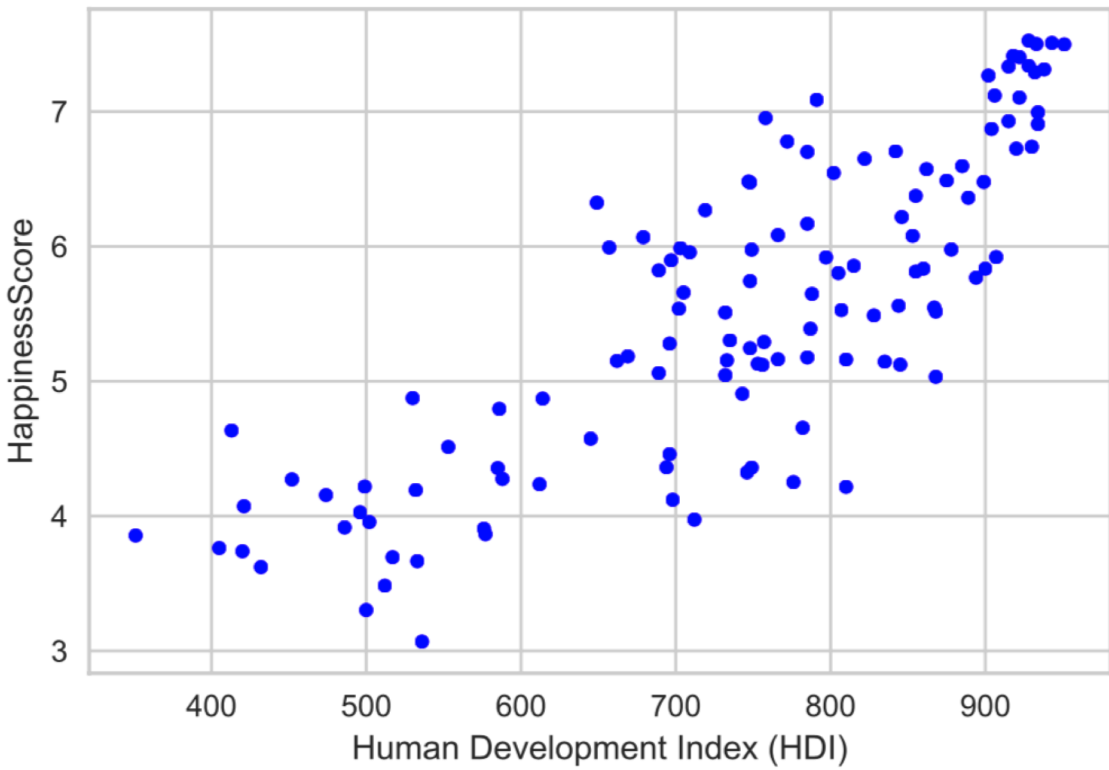
Linear regression is widely used



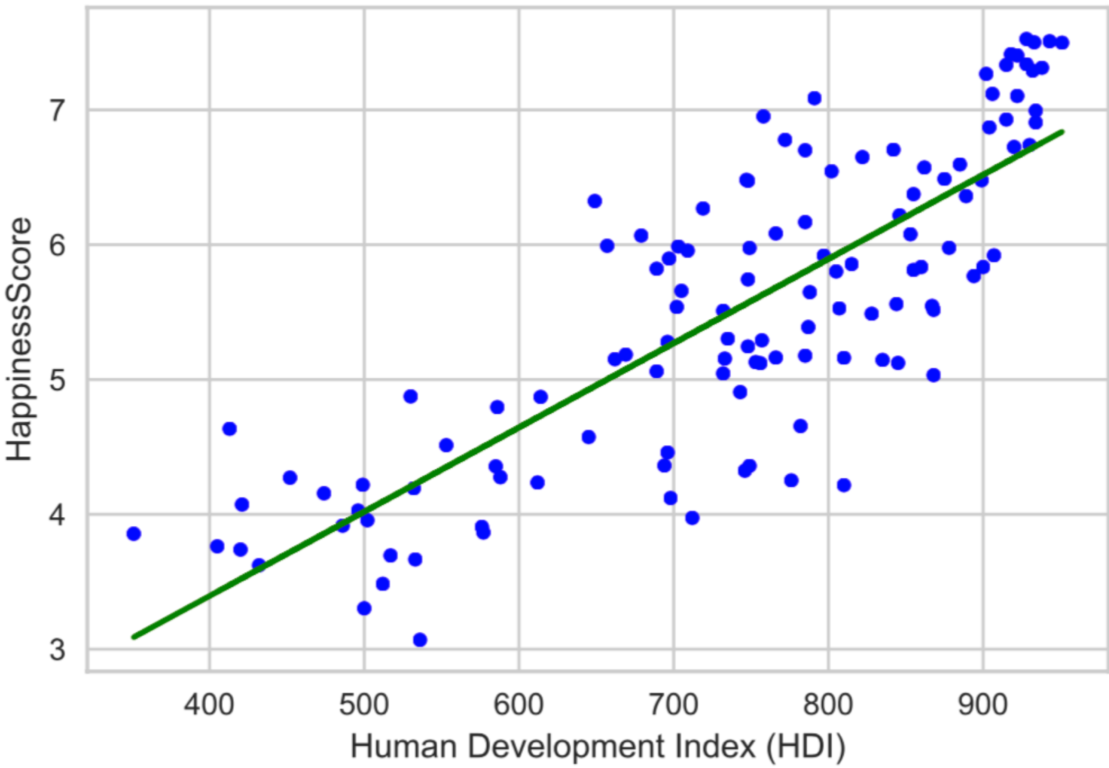
Source: https://www.kdnuggets.com/2019/04/top-data-science-machine-learning-methods-2018-2019.html

See also https://towardsdatascience.com/top-7-machine-learning-methods-that-every-data-scientist-must-know-84f5e5352ae1

Example: Does development of a country impacts citizen's happiness ?



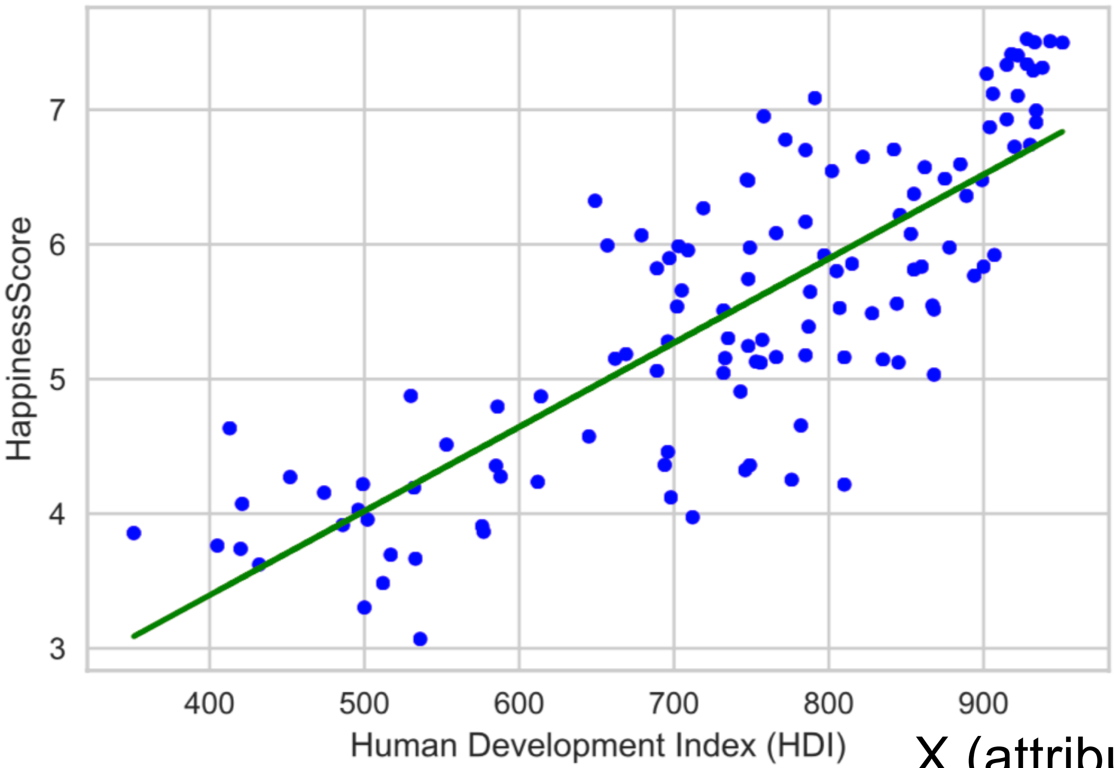
Example: Does development of a country impacts citizen's happiness ?



HappinessScore = 0.0062 HDI + 0.89 + *noise*

Example: Does development of a country impacts citizen's happiness ?

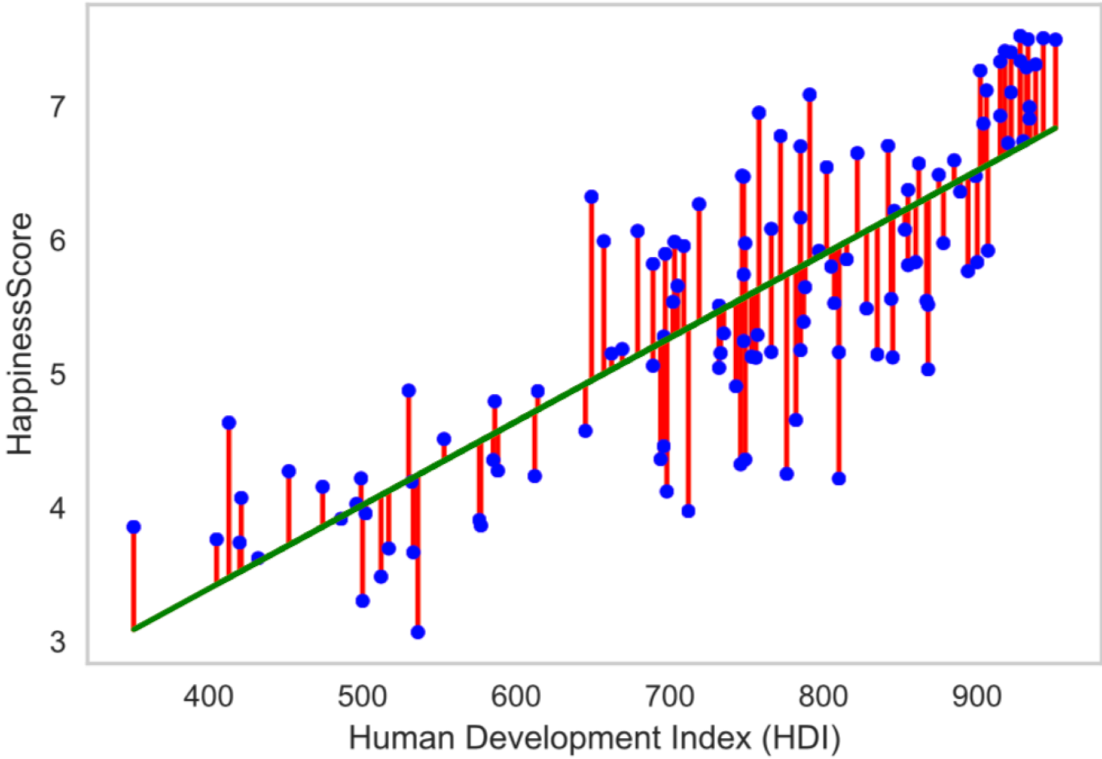
Y
(target)



X (attribute)

$$Y = 0.0062\,X + 0.89 + \varepsilon$$

We want to minimize the error…



$$Y = 0.0062\,X + 0.89 + \varepsilon$$

error term

Error terms in the 2-dimensional case $X = \{(x_1, x_2)\}$

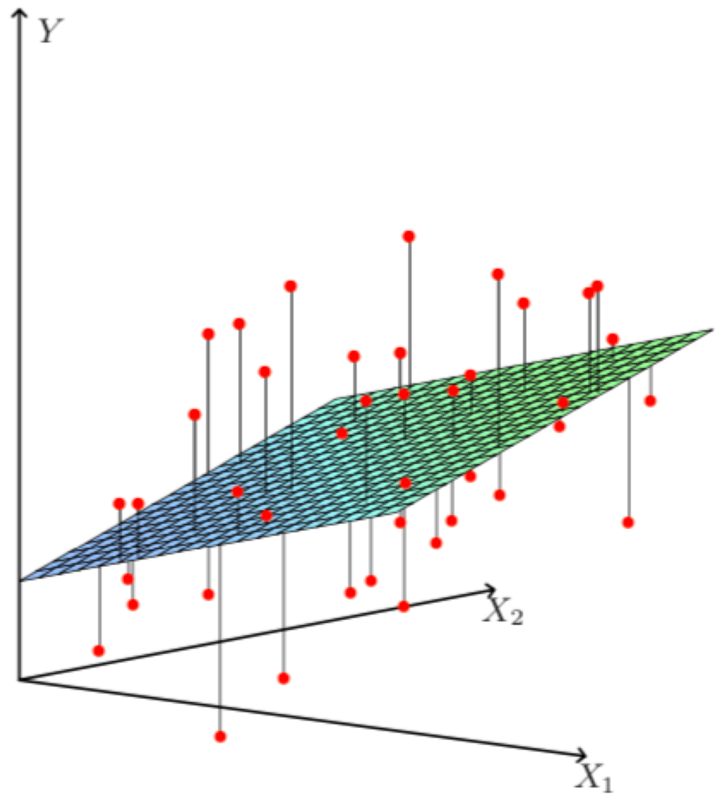


Illustration: The Elements of Statistical Learning

Linear regression models

Weight	Color	Seeds	Price
150	80	8	2.5
200	112	6	3.1
170	120	8	2.9
210	105	7	3.6
180	130	9	2.4

attributes x_i

target y

Predict (estimate) a real value *y*, from an input vector X

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

Note: The model is linear in its **parameters**, which means: $f(X, \alpha + \beta) = f(X, \alpha) + f(X, \beta)$

Linear regression models and least square

Error L2: $$\begin{aligned}\mathrm{RSS}(\beta) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2\end{aligned}$$

$\mathrm{RSS}(\beta)$ is a quadratic function, we find the minimum by taking the derivatives wrt β

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}.$$

Least Square Linear regression

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

$$\mathrm{RSS}(\beta) \quad = \quad \sum_{i=1}^{N} (y_i - f(x_i))^2$$

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Least Square Linear regression

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

$$\mathrm{RSS}(\beta) \quad = \quad \sum_{i=1}^{N} (y_i - f(x_i))^2$$

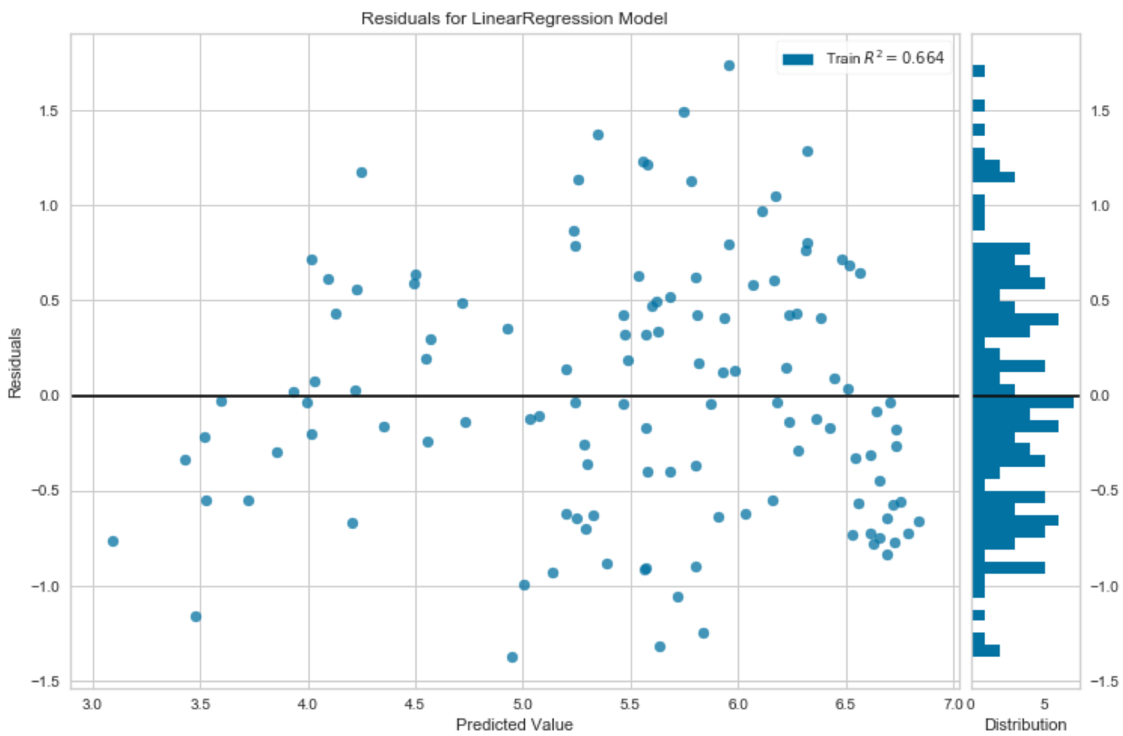
$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Implicit assumptions:

- The mean of the probability distribution of the error (RSS) is 0.
- The variance of the error is constant for all values of the predictor X.
- The probability distribution of the error term is normal.
- The values of the error term associated with any two observed values of y are independent. That is, the value of the error term associated with one value of Y has no effect on any of the values of the error associated with any other Y value.

Assessing the Quality of Linear Regression

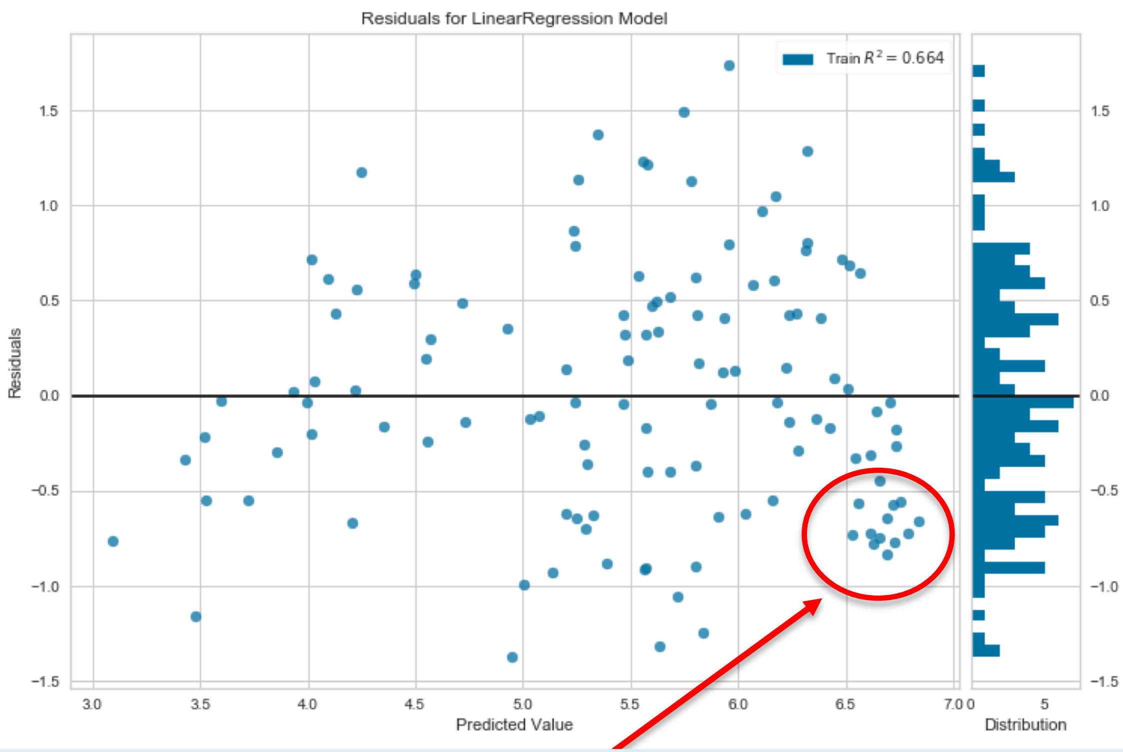
The **residuals plot** shows the difference between residuals (errors) on the vertical axis and the dependent variable on the horizontal axis, allowing you to detect regions within the target that may be susceptible to more or less error.



Plot easily done using scikit-learn and Yellowbrik, see https://www.scikit-yb.org/en/latest/api/regressor/residuals.html

Assessing the Quality of Linear Regression

The **residuals plot** shows the difference between residuals (errors) on the vertical axis and the dependent variable on the horizontal axis, allowing you to detect regions within the target that may be susceptible to more or less error.



Check *Homoscedasticity* (constant variance): variance of the errors should be constant with respect to the predicting variables or the response and ***Normality*** assumptions

In our example, the distribution of residuals is not really **normal**
=> This may reflect the non stationarity of the data
(see the cluster of rich countries with higher HDI and Happiness)

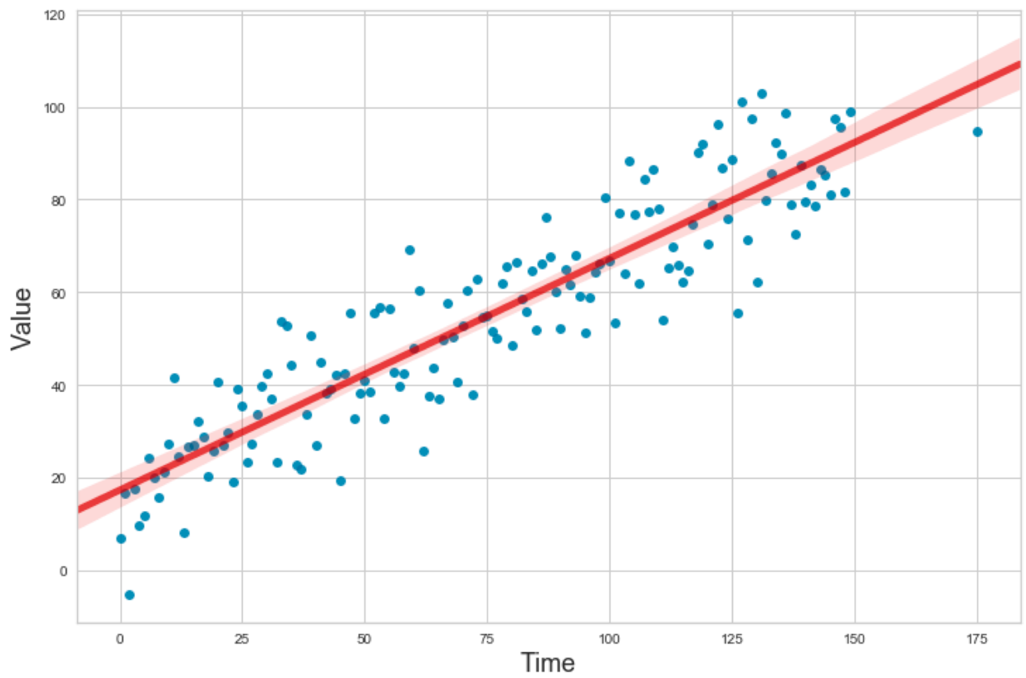
When dependency is not linear ?

- We can generate coordinates
  basis expansion, eg polynomials
  from (x_1, x_2) generate $(x_1, x_2, x_1^2, x_2^2, x_1 x_2)$
  and fit the model $f(X) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$
  (with 5 dimensions instead of 2)
  => See features engineering, ridge regression and lasso.

- Or use a generalized linear model (GLM)

Time series prediction with linear regression

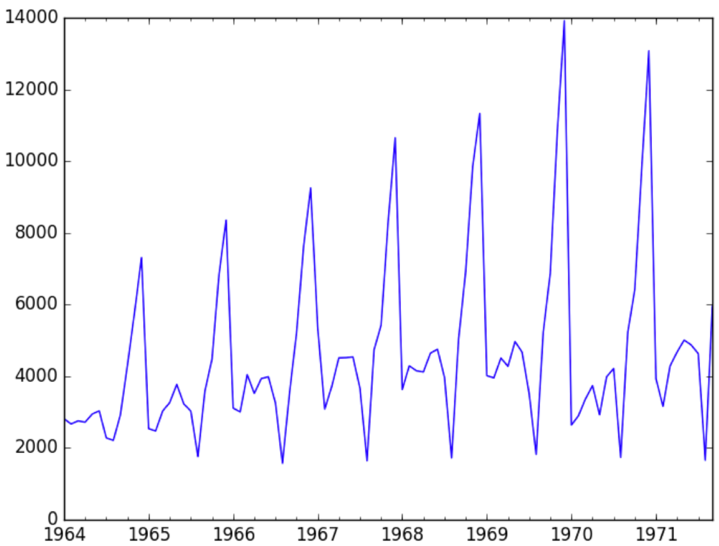
- The linear regression model can be used to make predictions :

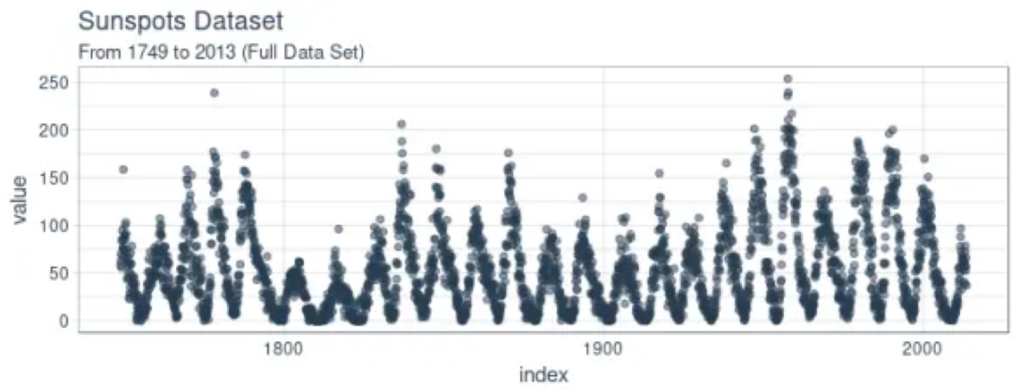


$$\hat{y}(t) = f(t) = \beta_0 + \beta_1 t$$

Time series prediction with linear regression

- But most series are not linear in time:





Time series exhibit **trends** and **seasonality**

Time series prediction with autoregression

- Model y(t) as a function of recent past values $\hat{y}(t) = \beta_0 + \beta_1\, y(t-1) + \beta_2\, y(t-2) \ldots$

$$\hat{y}(t) = \beta_0 + \sum_{\delta=1}^{H} \beta_\delta\, y(t-\delta)$$

Look for parameters minimizing the L2 error.

See also: ARIMA models.

Conclusion

We presented the basic **linear regression**, which is a convenient model used during preprocessing or as a baseline.

→ you should always try a simple linear regression before applying more complex models

- Next: non linear regression, regularization, Ridge regression and Lasso.

Quizz

1. What is the difference between a classification model and a regression ?
2. What kind of data if necessary for supervised regression ?
3. What is the error criteria optimized by standard linear regression ?
4. To what indicators should we look to assess the quality of a linear regression fit ?
5. How can we use linear regression models to predict the next values of a time series ?

References

Books

- T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Springer, 2017 https://web.stanford.edu/~hastie/ElemStatLearn/

Tutorials

- Introduction for beginners: https://www.surveygizmo.com/resources/blog/regression-analysis/
- Quality checks: https://www.kdnuggets.com/2019/07/check-quality-regression-model-python.html
- Scikit-learn, software tools & tutorials:
    - A Beginner's Guide to Linear Regression in Python with Scikit-Learn https://www.kdnuggets.com/2019/03/beginners-guide-linear-regression-python-scikit-learn.html
    - Visualizing linear relationships with Seaborn https://seaborn.pydata.org/tutorial/regression.html
    - Residual plots https://www.scikit-yb.org/en/latest/api/regressor/residuals.html
    - Generalized Linear Models https://scikit-learn.org/stable/modules/linear_model.html
    - Forecasting Time Series : https://pythondata.com/forecasting-time-series-autoregression and https://towardsdatascience.com/time-series-analysis-in-python-an-introduction-70d5a5b1d52a)