Brief Data Science Concepts,
Overview of Statistical Learning,
by Matthew Balogh,
Issue 1

Contents

1 Preface

Brief Data Science Concepts encloses my overview of certain topics and my practical experiments in the field of *Data Science*, as its name implies, in a short form.

In this first *issue*, I overviewed the concept of *Statistical Learning* guided by *An Introduction to Statistical Learning with Applications in Python* written by James et al. (2020). The objective was to revisit the fundamental aspects of statistical and machine learning techniques and reconstruct datasets and visuals presented in the referenced work.

As a result, from the theoretical overview part of this project, I revisited the concepts of predictor and response variables, observations and their mathematical and tabular representations. Moreover, clarified the differences between prediction and inference, reducible and irreducible errors, true relationship and an estimator.

From a technical point of view, I gained experience in synthetic data generation, training various estimator models, along with creating meaningful and easily interpretable visualizations with the tools of matplotlib, plotly, and the packages of pandas, numpy in Python.

The associated Python notebook is available here.

And my **Linked In** profile can be found here.

2 Motivating Scenario

Given two independent variables education and seniority both in a non-linear relationship with the target variable of income as shown in Figure 1. That is, the two variables education and seniority individually contribute to the final income of an individual as depicted in Figure 2. The relationship itself is synthetically generated exhibiting income over years of education and an arbitrary scoring scale associated with seniority. A collection of 30 data points are observed from this variable space and is referred to as the **Income** dataset. Observations represent the scenario where we may already have or acquire education and seniority characteristics of 30 individuals along with their income. Then, we aim to use these records to infer about the underlying relationship and make predictions for new observations related to their target variable. In a very simple case, it might happen that the phenomenon to be observed depends on only one predictor. A dataset named **IncomeSimple** will be used for this scenario. The datasets are visualized in Figure 3 and 4.

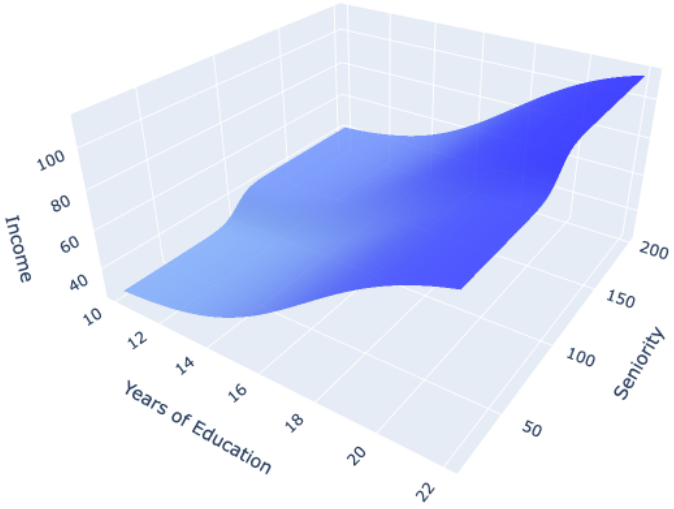

Figure 1: The synthetic relationship of income over education and seniority. It is an additive relationship incorporating the two non-linear relationships from Figure 2.

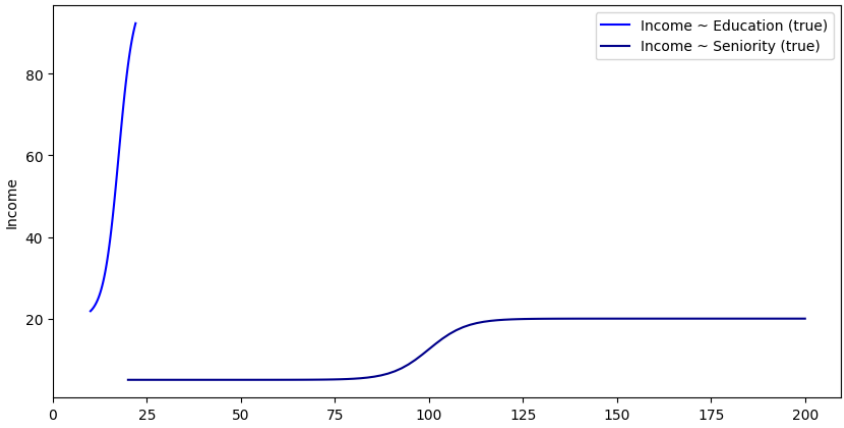

Figure 2: Individual contribution effects of education and seniority on income.

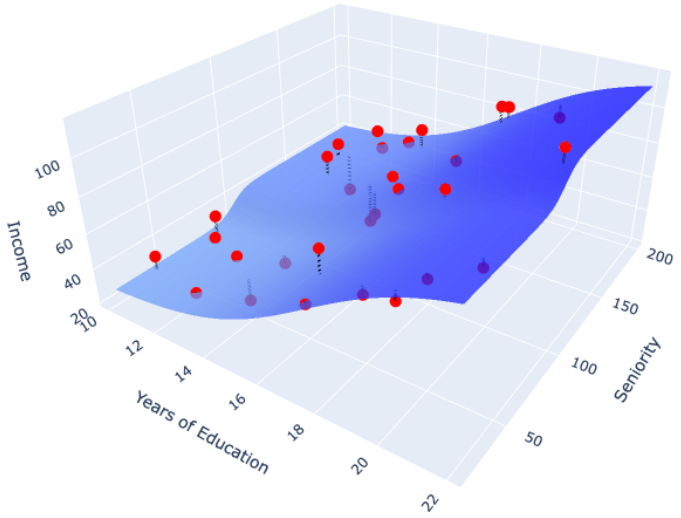

Figure 3: A collection of observations drawn from a synthetic relationship of income over education and seniority, to be referred as the **Income** dataset. Observations are marked as red points, while their deviations from the true relationship function—implied by noise—are displayed with vertical black lines. The observations are generated to be diluted with zero-mean noise.

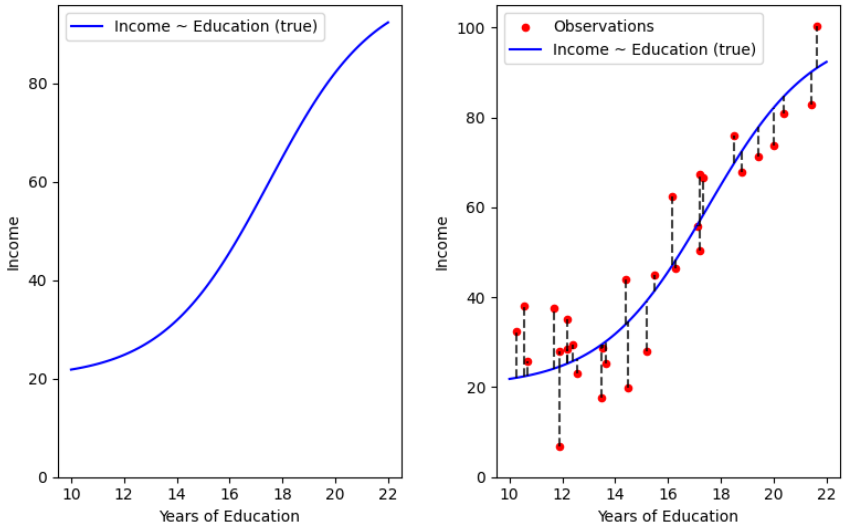

Figure 4: The synthetic relationship of income over education. A collection of observations drawn from a synthetic relationship to be referred as the **IncomeSimple** dataset is displayed in the right plot. The observations are generated to be diluted with zero-mean noise, and there are no additional predictors contributing to the response. The blue curve shows the true underlying relationship, and residuals of the observations are displayed with vertical black lines.

3 Statistical Learning

3.1 Predictors and Response

In the data collection and learning process, there are input and output variables. As James et al. (2020) describe it, an input variable is called *predictor*, *feature*, *independent variable*, or simply *variable*, while an output variable goes by the names of *target variable*, *response*, or *dependent variable*. In the introduced scenario, considering the **Income** dataset, education and seniority are predictors, while income is the response variable.

3.2 Observations

A collection of data points, the observations are used as training data to teach a specific method to estimate f, the true underlying relationship between the predictors and response in question. The observations are denoted as follows.

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \ , \tag{1}$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, that is, each encloses all the p predictor values for the ith observation. As for an example, observations from the **Income** dataset are depicted in Figure 5 in a tabular form, where each row represents a single observation, while columns stand for predictors and the response variable. Figure 6 depicts observations from the **IncomeSimple** dataset.

	education	seniority	income
0	14.494481	129.358073	31.874537
1	21.408572	50.694342	90.341105
2	18.783927	31.709287	73.070548
3	17.183902	190.799397	86.519782
4	11.872224	193.813766	48.237728

Figure 5: A sample drawn from the **Income** dataset represented in tabular form. Rows represent individual observations, while columns stand for variables.

	education	income
0	14.494481	19.775928
1	21.408572	82.874615
2	18.783927	67.809389
3	17.183902	67.416804
4	11.872224	27.962446

Figure 6: A sample drawn from the **IncomeSimple** dataset represented in tabular form. Rows represent individual observations, while columns stand for variables.

3.3 Prediction and Inference

Estimating the true relationship f is performed in order to make *inference* or *prediction*. According to the authors (James et al. 2020), during a prediction, the estimator function $\hat{f}$ is treated as a black box and the interest is in obtaining a response for the given predictors corresponding to unseen observations. During inference, however, the exact form of of the estimator, the association between the predictors and the response variable is of interest. For example, we may wish to establish an estimator to predict the income (y_{i+1}) of an unseen observation with predictor values $x_{i+1} = (13, 100)^T$, for education and seniority, respectively. Or we may be interested in finding out what curve the data points take on, and infer about the underlying relationship and make conclusions based on that.

3.4 Relationship

Given a set of p predictors, $X = (X_1, X_2, \ldots, X_p)$ and a response variable Y, it is assumed that there is some relationship between the predictors and the response, which is denoted as follows.

$$Y = f(X) + \epsilon \ , \tag{2}$$

where "f is some fixed but unknown function of $X_1, X_2, \ldots, X_p$ and ϵ is a random error term, which is independent of X and has mean zero (James et al. 2020)". An example is shown in Figure 7, which suggests that there is a relationship between income and the years of education considering the **IncomeSimple** dataset. Figure 8 exhibits a similar scenario, yet with an additional predictor seniority included, that is considering the **Income** dataset.

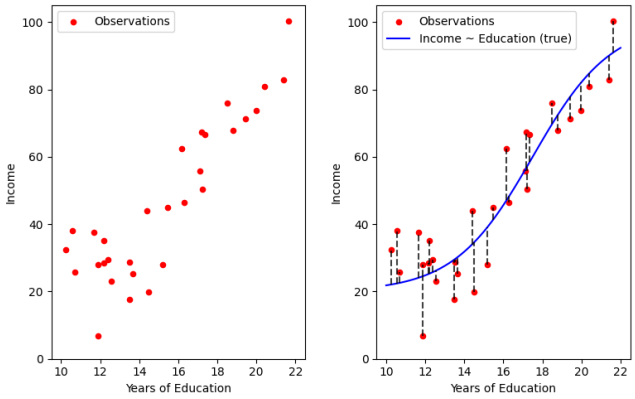

Figure 7: A collection of observations (**IncomeSimple**) showing the trace of a relationship between education and income. The true relationship is highlighted in the right plot.

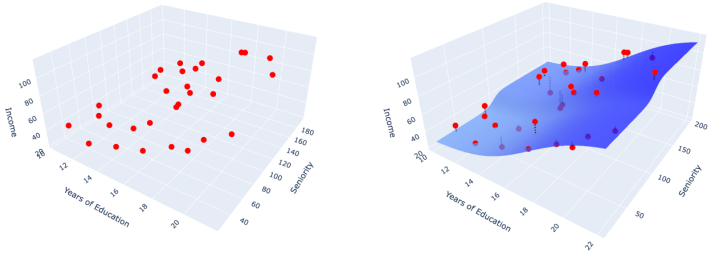

Figure 8: A collection of observations (**Income**) showing the trace of a relationship between education, seniority, and income. The true relationship is highlighted in the right plot.

3.5 Estimation

The objective is to estimate f by establishing an estimator $\hat{f}$ by employing various statistical learning approaches such that it approximates the given observations. That is:

$$Y \approx \hat{f}(X) \ , \tag{3}$$

for any (X, Y) observation.

3.6 Reducible and Irreducible Error

It is explained in *An Introduction to Statistical Learning with Applications in Python*, that one can distinguish between reducible and irreducible error when estimating a relationship between variables. It is shown by the authors that if one could establish an estimator equivalent to the true relationship, then $\hat{f}(X) = f(X) = \hat{Y}$ would not result in the observed values Y, since Y incorporates the error term ϵ as $Y = \hat{f}(X) + \epsilon$. Conversely, if one would model the observations Y perfectly, the model would fail to capture the true relationship due to the noise incorporated in the model. Since the true relationship is often challenging to be reproduced by the estimator learned from the limited observations, $\hat{f}$ will differ from f, introducing another error term besides ϵ. In other words, the error implied by the differences between the true relationship function f and the estimator $\hat{f}$ can be reduced by using the appropriate methods, hence it's called reducible error, while the error term ϵ caused by the noise of the observations are referred to as the irreducible error, which as the authors highlight, cannot be predicted using X. The authors also emphasize that the quantity ϵ may represent more than just the noise. For example, unmeasured variables, or the effect of ignored variables may be obscured and appear in ϵ. Moreover, factors such as a patient's response based on their well-being at a moment may affect their reaction, therefore the variance in the response variable (James et al. 2020). Figure 9 depicts how the true relationship function can be estimated with a finer level of detail in order to reduce the difference between f and $\hat{f}$ by following more flexible approaches, in the referenced case non-linear polynomial fitting.

3.7 Parametric Methods

The authors introduce the *parametric* method as a model-based approach, where an assumption about the functional form of f is made. This assumption reduces the complexity of the estimation problem. After a model has been selected, a procedure is performed on the training data that estimates the parameters of the chosen model. For example, we may assume that

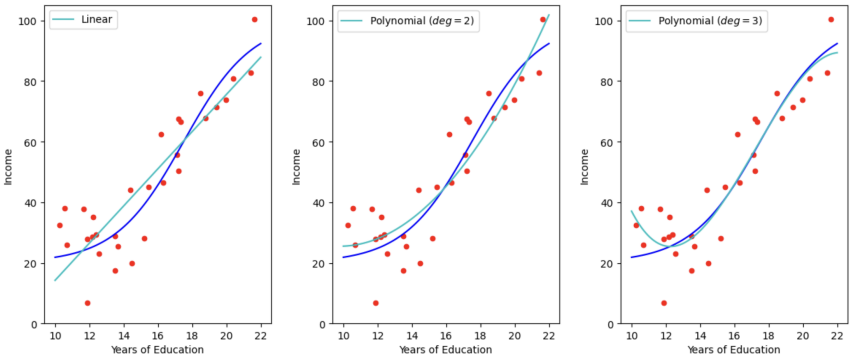

Figure 9: Data points (red dots) of **IncomeSimple** and estimator functions (cyan line and curves) of the true relationship (blue curve) with linear, polynomial of degree 2, polynomial of degree 3 models, respectively. Linear model is too simple to capture the true relationship, while polynomials are able to better fit it. The predictor education is normalized.

the observations follow a linear or non-linear, polynomial relationship, and employ least squares approach to obtain an estimator as it was performed in the plots of Figure 9. The key thing here is that estimating a set of parameters such as $\beta_0, \beta_1, \ldots, \beta_k$ is much easier than estimating an arbitrary shaped function f. A too simplistic model, however, may fit the training data poorly, while a more complex model—aside from increasing the estimation problem—may capture the noise of the training data. While Figure 9 relates to **IncomeSimple**, Figure 10 showcases parametric approaches to estimate the relationship of the **Income** variables utilizing linear along with different polynomial and spline models.

3.8 Underfit and Overfit

A too simple model provides a smooth estimation which not only diminishes the effect of noise but overlooks the main characteristics and patterns of the phenomenon being observed, leading to *underfit*. On the other hand, however, a complex method may fit the training data too-precisely, in a way that it not only takes finer details of the training data into account but so does the noise. It learns patterns that are not present in the data in general, leading to *overfit*. Underfitting and overfitting models are displayed in Figure 11 for the **IncomeSimple** dataset.

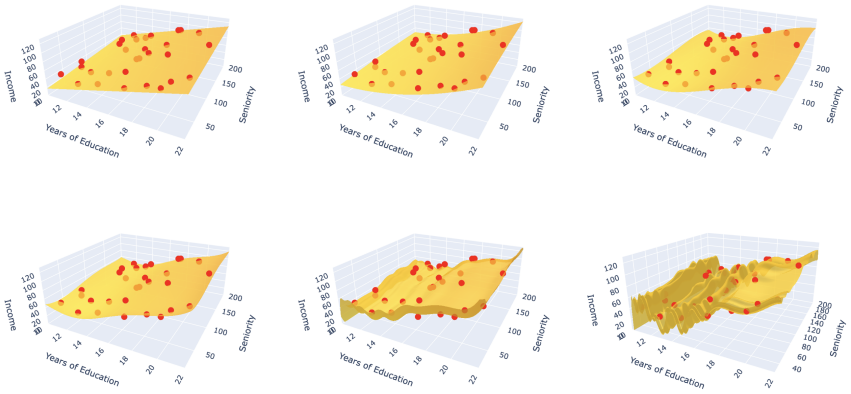

Figure 10: Parametric approaches for estimating the relationship of the **Income** dataset. The selected models predictor-wise are: *linear*, *second-degree polynomial*, and *third-degree polynomial*, respectively in the first row, and *spline* models with different hyperparameters in the second row.

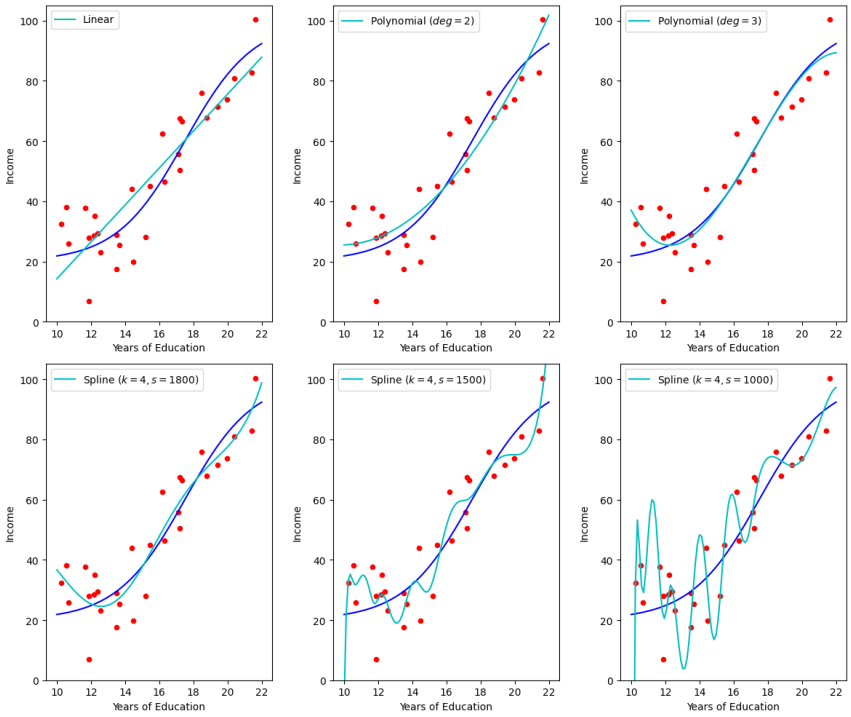

Figure 11: The phenomena of *underfit* and *overfit*. The result of going from a too simple to a too complex model, considering the **IncomeSimple** dataset.

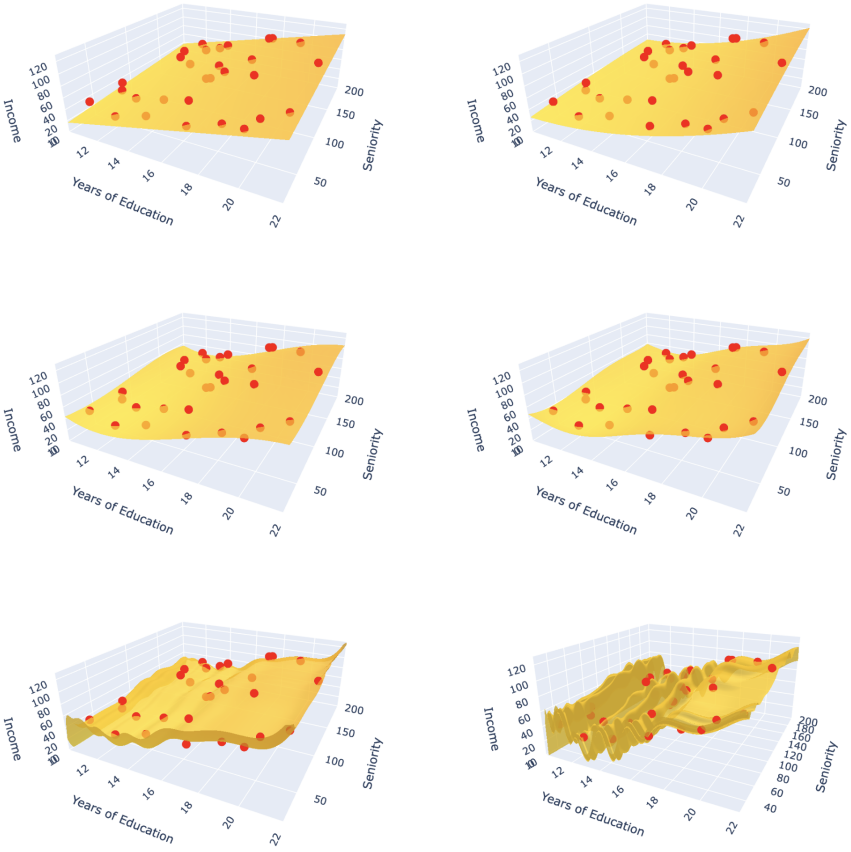

Figure 12: Detailed version of Figure 10; Parametric approaches for estimating the relationship of the **Income** dataset. The selected models predictor-wise are: *linear*, *second-degree polynomial*, and *third-degree polynomial*, respectively in the first row, and *spline* models with different hyperparameters in the second row.

References

James, Gareth et al. (2020). *An Introduction to Statistical Learning with Applications in Python*. 1st. New York: Springer.