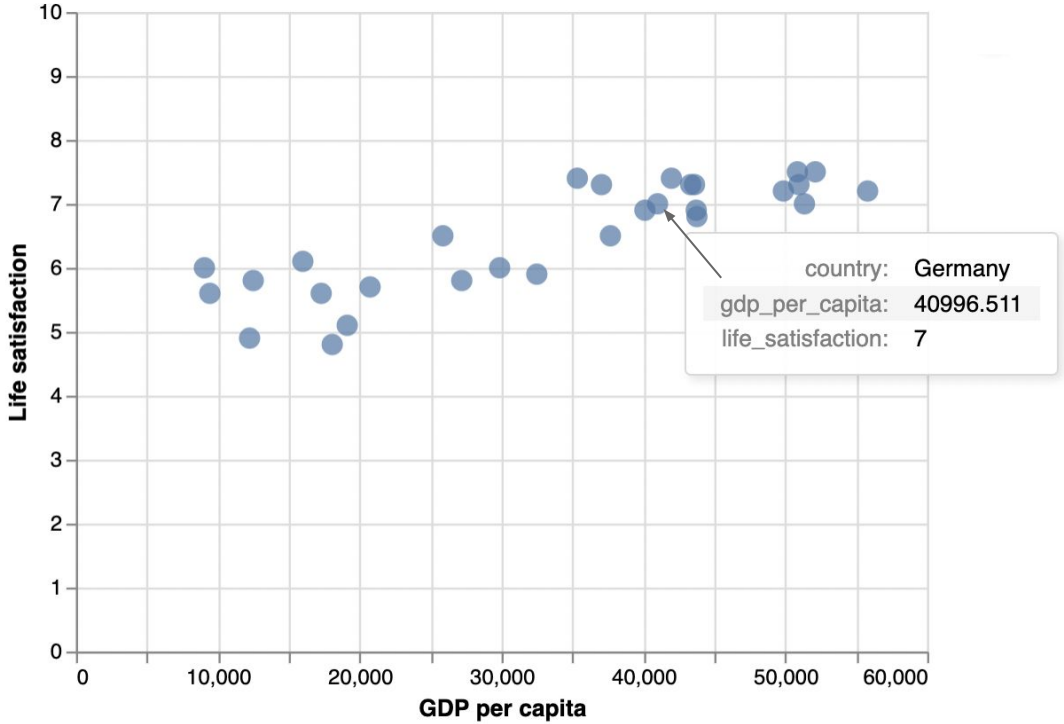
Regression models

Does money make people happier?

Prof. Dr. Jan Kirenz

Does money make people happier?



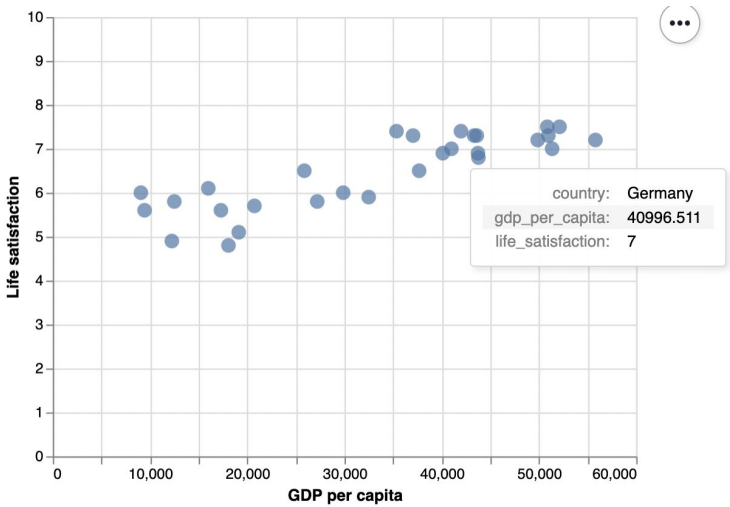
Source: Géron (2019)

Question:

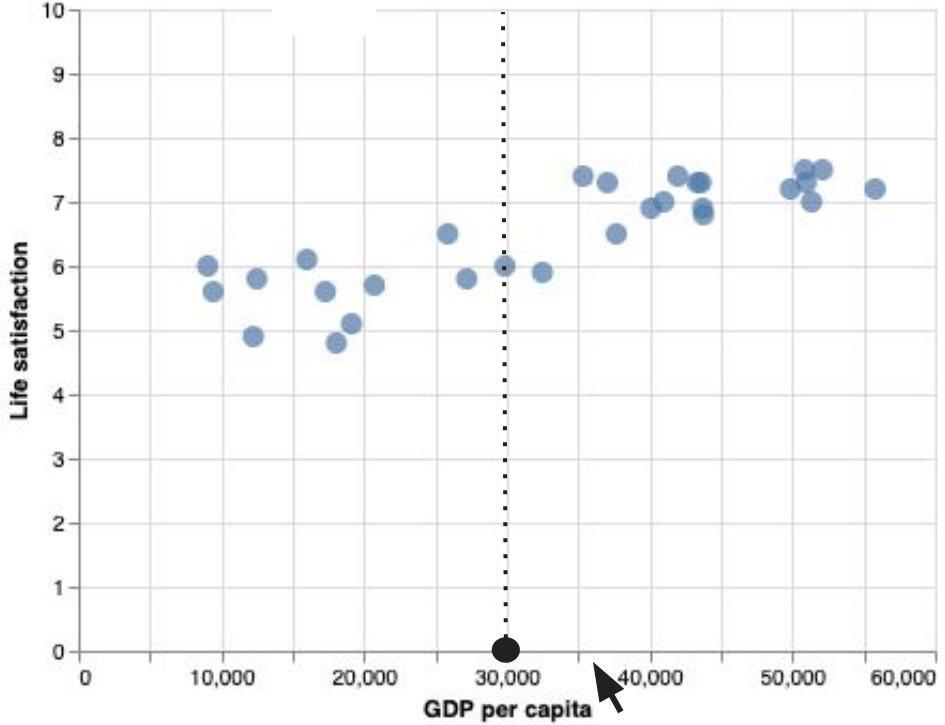
1. **Data exploration** gives indication for a trend that is:
   a. Positive or negative?
   b. Linear or non-linear?

... but be careful, the data is noisy (i.e., partly random)

2. **Model (type) selection:**
   a. Regression or classification?
   b. Linear or non-linear?



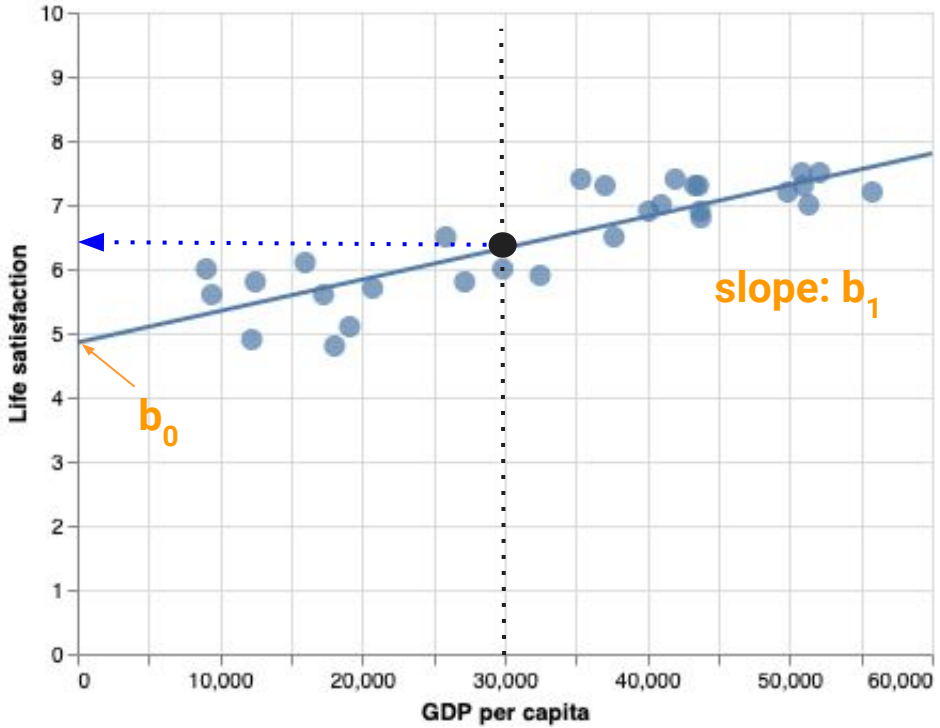
How to predict the life satisfaction of new country with a GDP per capita of 30000?



A simple linear regression model

Outcome = **Model**(GDP)

Life Satisfaction = $b_0 + b_1 \times$ GDP



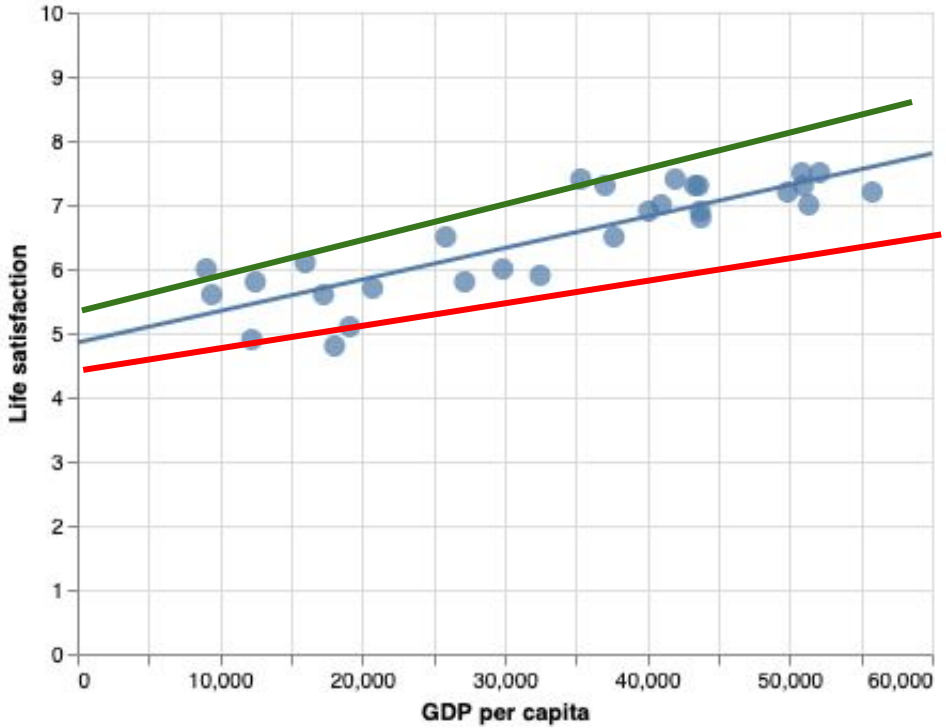
Source: Géron (2019)

How to select the best model?

A few possible linear models with different parameters

Outcome = **Model**(GDP)

Life Satisfaction = $b_0 + b_1 \times$ GDP



Source: Géron (2019)

Model can refer to a

1. Type of model
   - e.g., Linear Regression
2. Fully specified model architecture
   - e.g., Linear Regression with one input and one output.
3. Final trained model
   - e.g. Linear Regression with one input and one output, using $\boldsymbol{\theta}_0 = 4.85$ $\boldsymbol{\theta}_1 = 4.91 \times 10^{-5}$

Source: Géron (2019)

Model selection includes

1. Choosing the type of model
    - *e.g., Linear Regression*
2. Fully specifying its architecture
    - *e.g., Linear Regression with one input and one output.*
3. Fitting (training) the model to find the model parameters that will make it best fit the training data
    - e.g. Linear Regression with one input and one output, using $b_0 = 4.85$ $b_1 = 4.91 \times 10^{-5}$

Source: Géron (2019)

A simple linear regression model

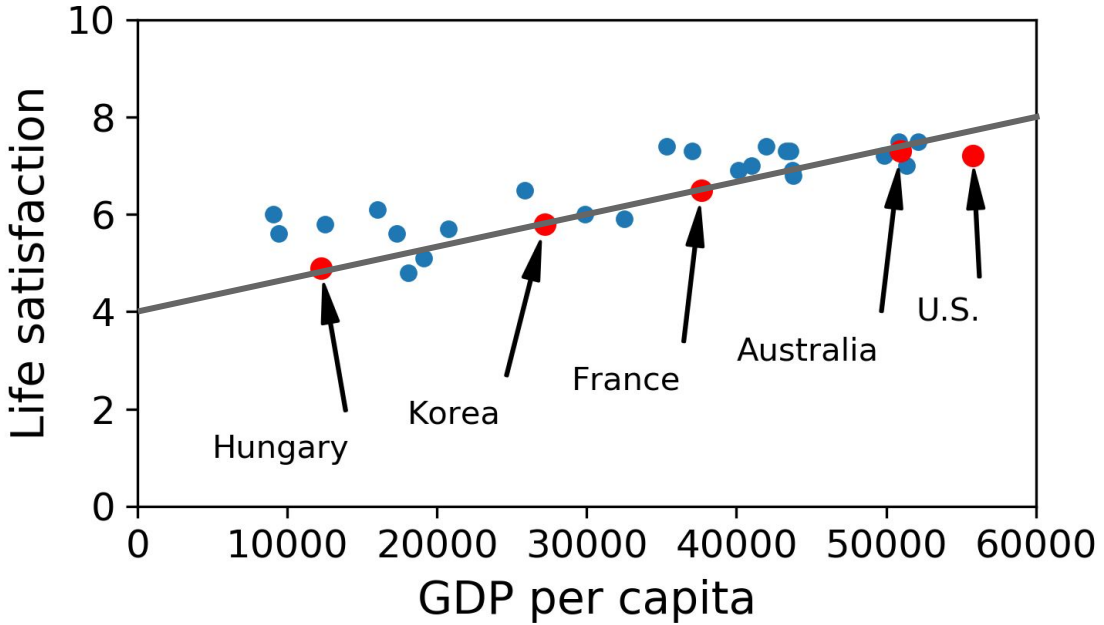
$$\hat{y}_i = \theta_0 + \theta_1 \times x_1$$

- $\hat{y}_i$ is the predicted output (life satisfaction).
- θ_0 is the bias (the y-intercept).
- θ_1 is the slope of our feature 1 (in machine learning often called weight of the feature)
- x_1 is our feature GDP (a known input).

All the same, just different notations:

$\hat{y}_i = b_0 + b_1 \times x_1$

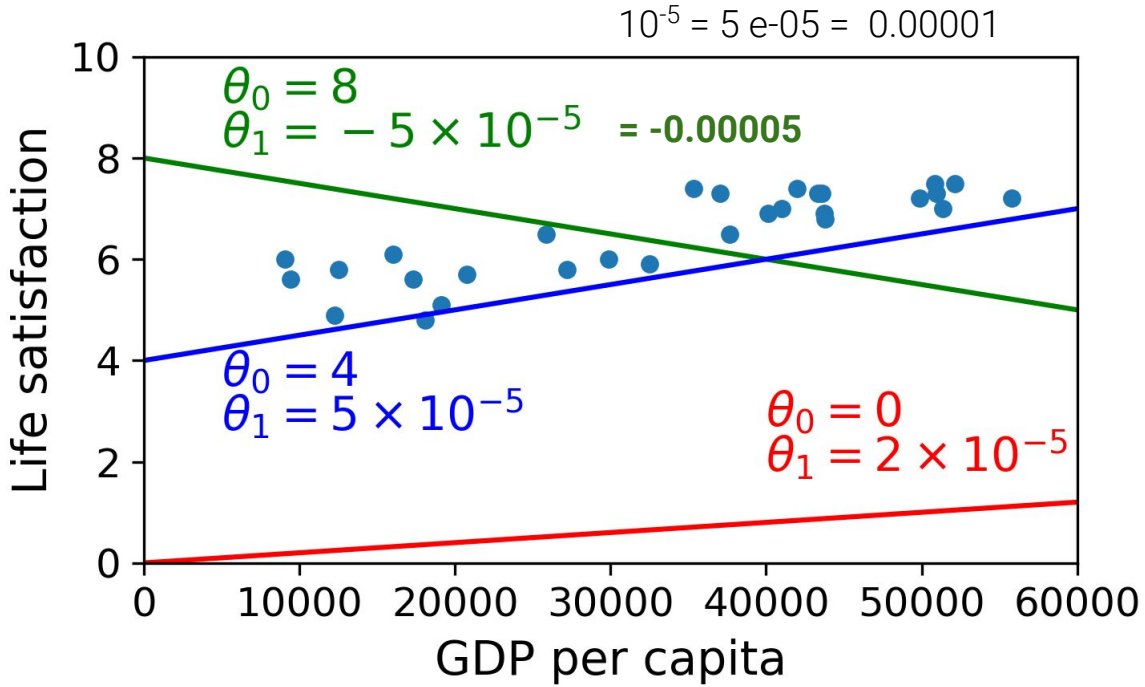
$\hat{y}_i = w_0 + w_1 \times x_1$



Source: Géron (2019)

How to select the best fitting model?

A few possible linear models with different parameters (θ = Theta)



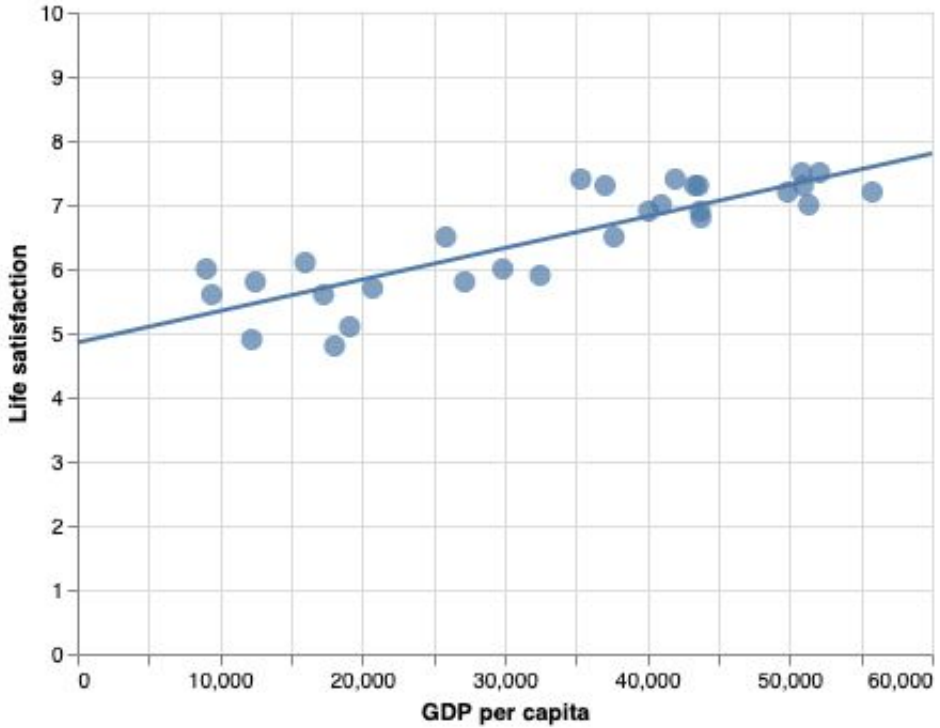
Source: Géron (2019)

Our **best fitting model**

We used the **mean squared error** to select the best model.

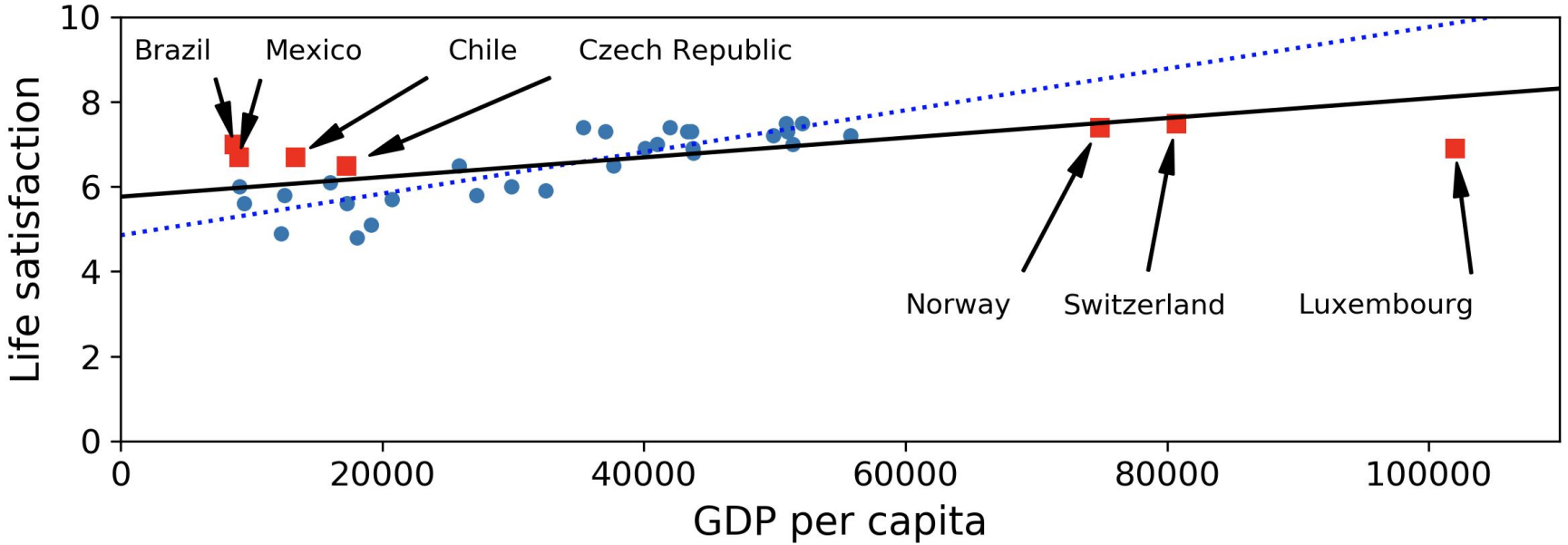
Life Satisfaction = $\mathbf{b_0}$ + $\mathbf{b_1}$×GDP

- b_0 = 4.85
- b_1 = 0.0000491 (4.91e-05)



Source: Géron (2019)

Our “old” best fitting model and more **representative data**



Source: Géron (2019)

We want our model to **generalize well**

That means data needs to be **representative**.

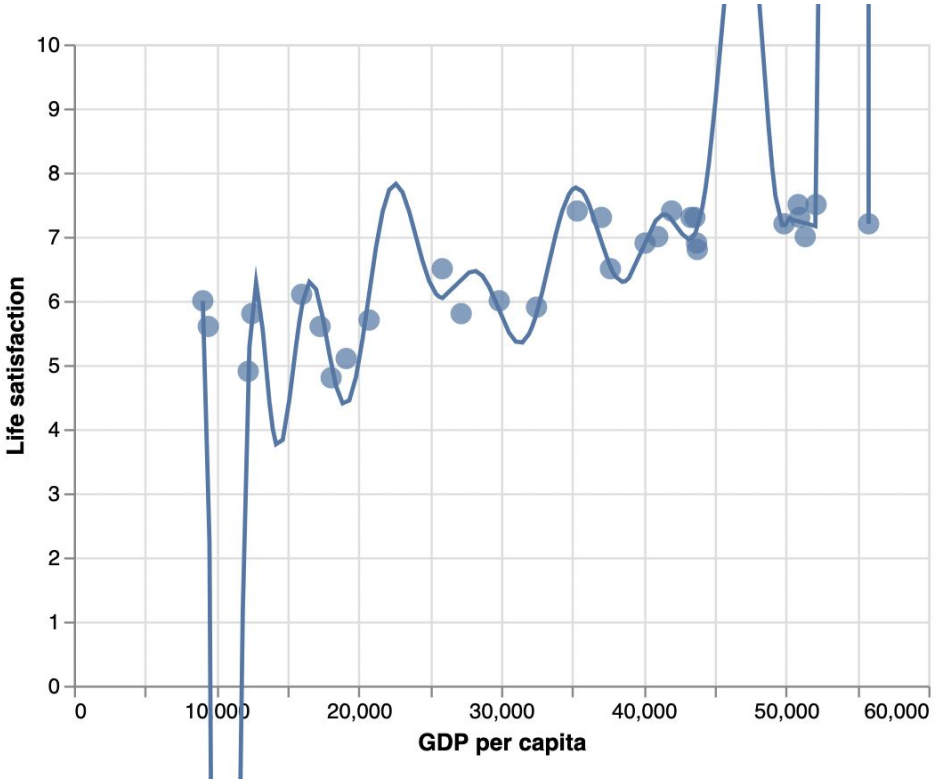
Possible data issues:

1. dataset to small: sampling noise

2. sampling method flawed: sampling bias

Source: Géron (2019)

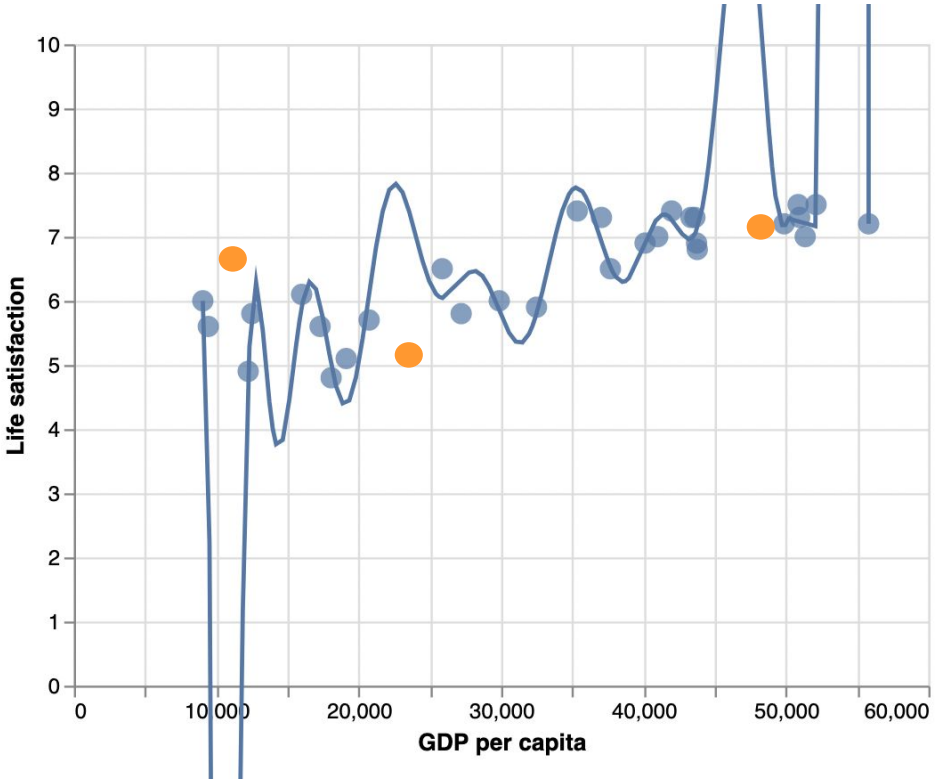
Can a model be “too good”?

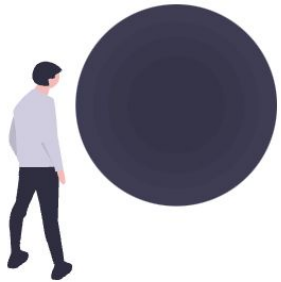


Our complex model makes almost no errors!

Source: Géron (2019)

How well does it predict **new data**?

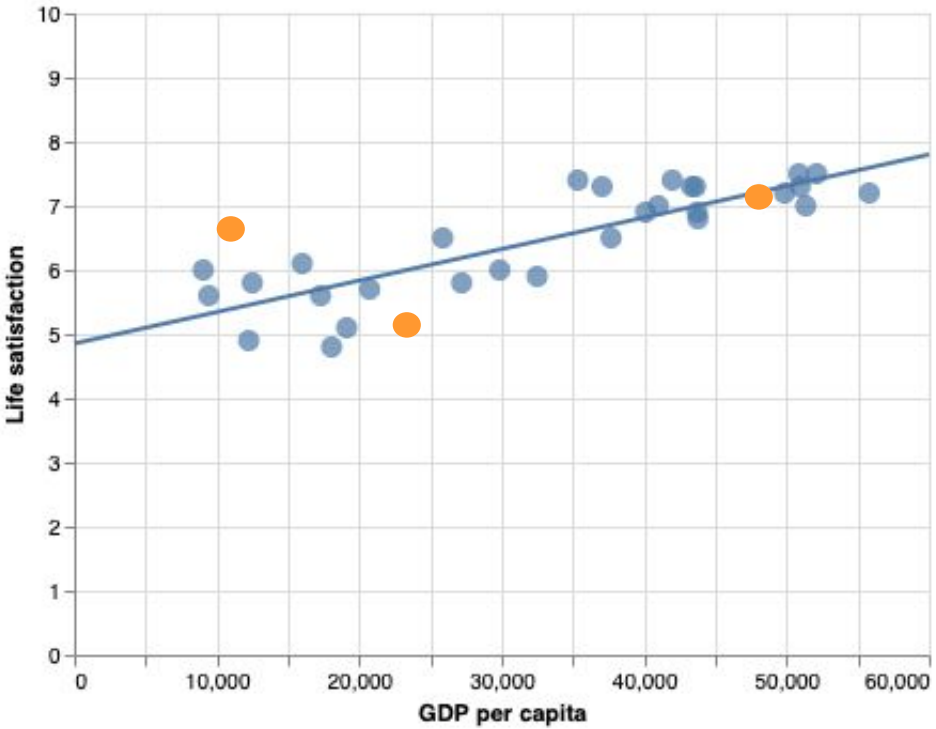




Oh no!

Source: Géron (2019)

How well does this simple model predict **new data**?

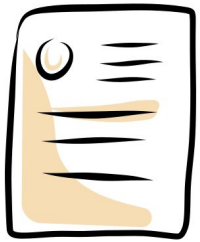


Source: Géron (2019)

We need to know how well our model predicts “new” data

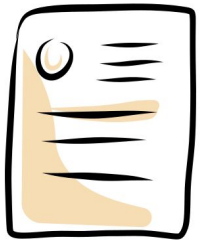
Source: Géron (2019)

Lab: Jupyter Notebook



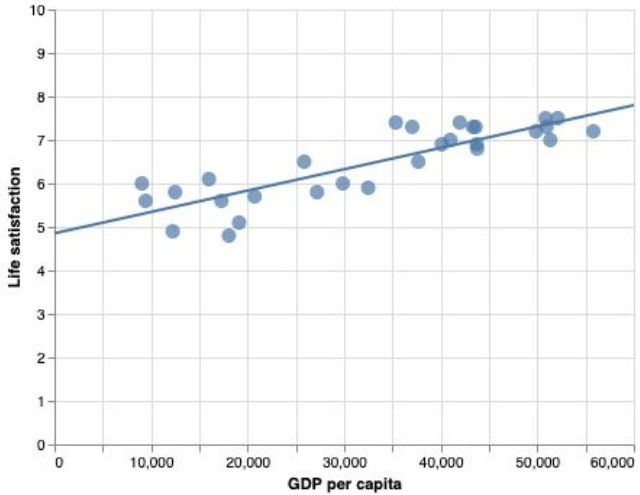
Build a regression model

Lab: Jupyter Notebook



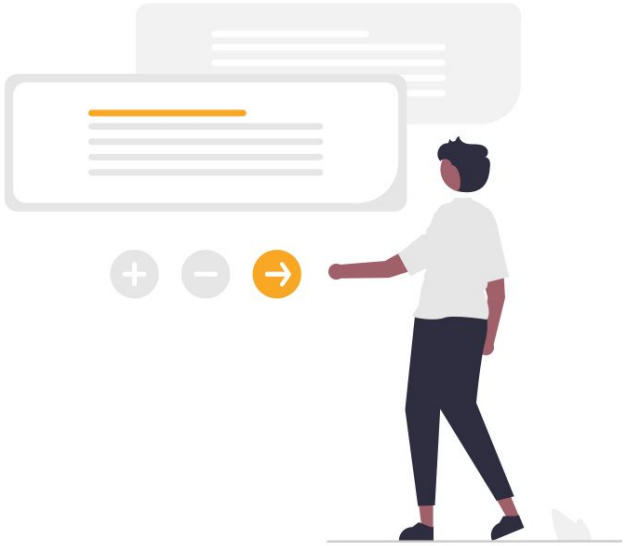
Use your model to predict the life satisfaction of Cypriots.

Cyprus's GDP per capita: $22587.



What’s next?

- Data splitting



Literature

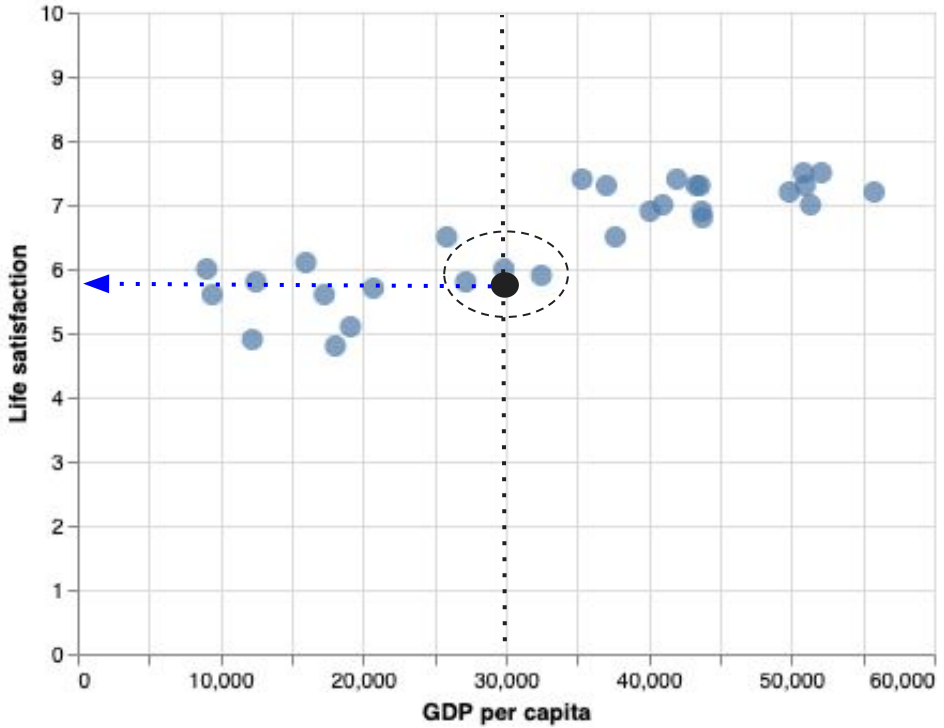
Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2 edition)*. O'Reilly Media.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). An introduction to statistical learning. New York: Springer.

Backup

K nearest neighbors model (here with k=3)

- Prediction by local interpolation of the 3 targets associated of the nearest neighbors in the data.
- We take the mean value of these 3 countries

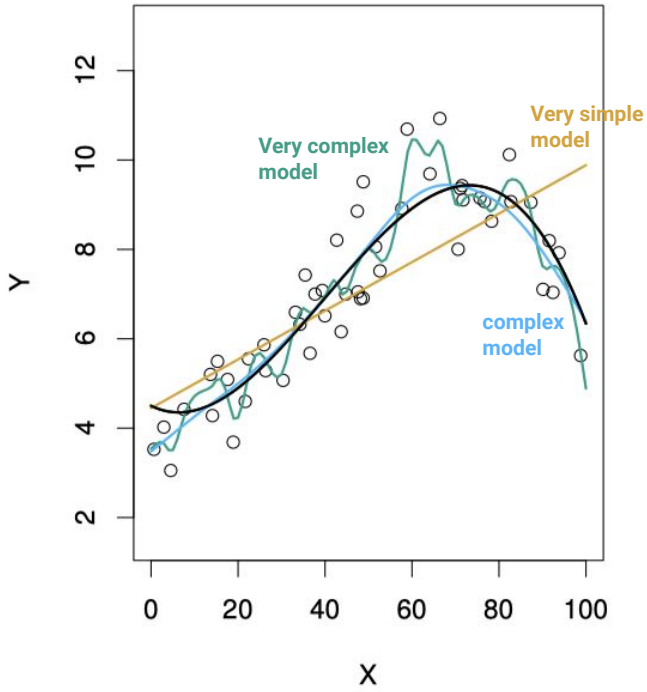


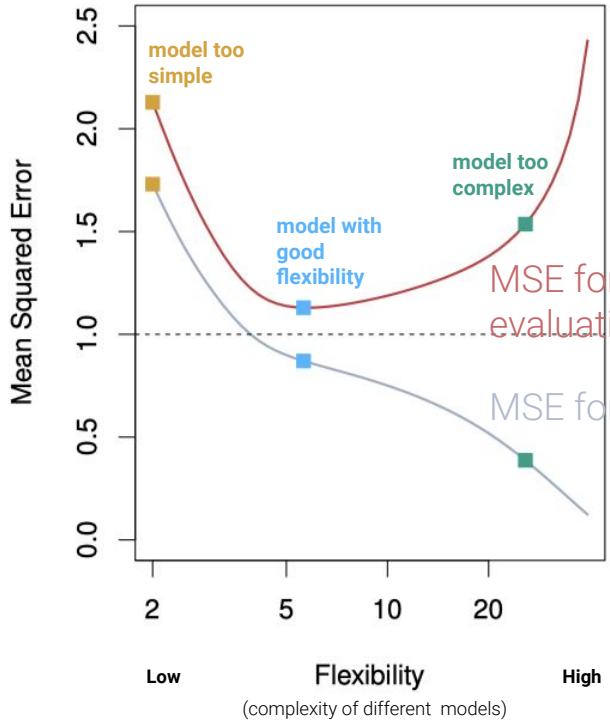
Bias-Variance trade-off

Underfitting ⟷ Overfitting

High bias
Low variance

Low bias
High variance





MSE for Test data (from evaluation set)

MSE for Training data

Source: James et. al. (2021)

Does money make people happier?

- Get the data at GitHub:

  GitHub

- Code in Colab:

  colab

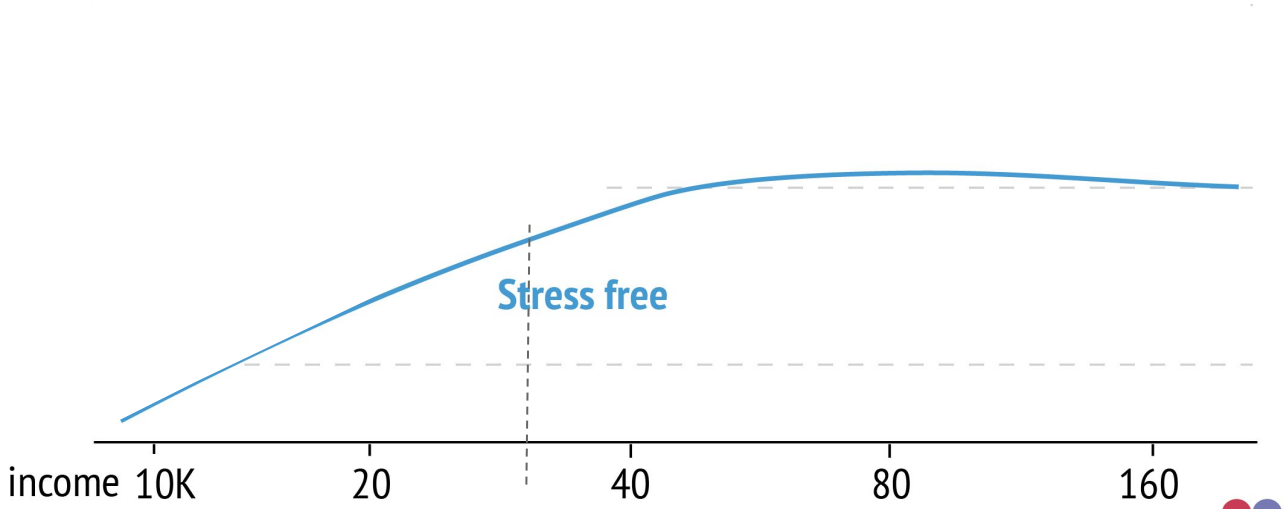
Country	GDP per capita	Life satisfaction
Hungary	12239.894	4.9
Korea	27195.197	5.8
France	37675.006	6.5
Australia	50961.865	7.3
United States	55805.204	7.2

Raw data:

OECD Better Life Index data: **Life satisfaction**

OECD.Stat

IMF: **Gross domestic product per capita**



Note: X axis scale is not linear.
Source: Kahneman and Deaton (2010)

CNBC

„Below an income of … $60,000 a year, people are unhappy, and they get progressively unhappier the poorer they get.

Above that, we get an absolutely flat line. … Money does not buy you experiential happiness, but lack of money certainly buys you misery."

Watch TED-talk:

TED

Money can buy happiness, but only to a point