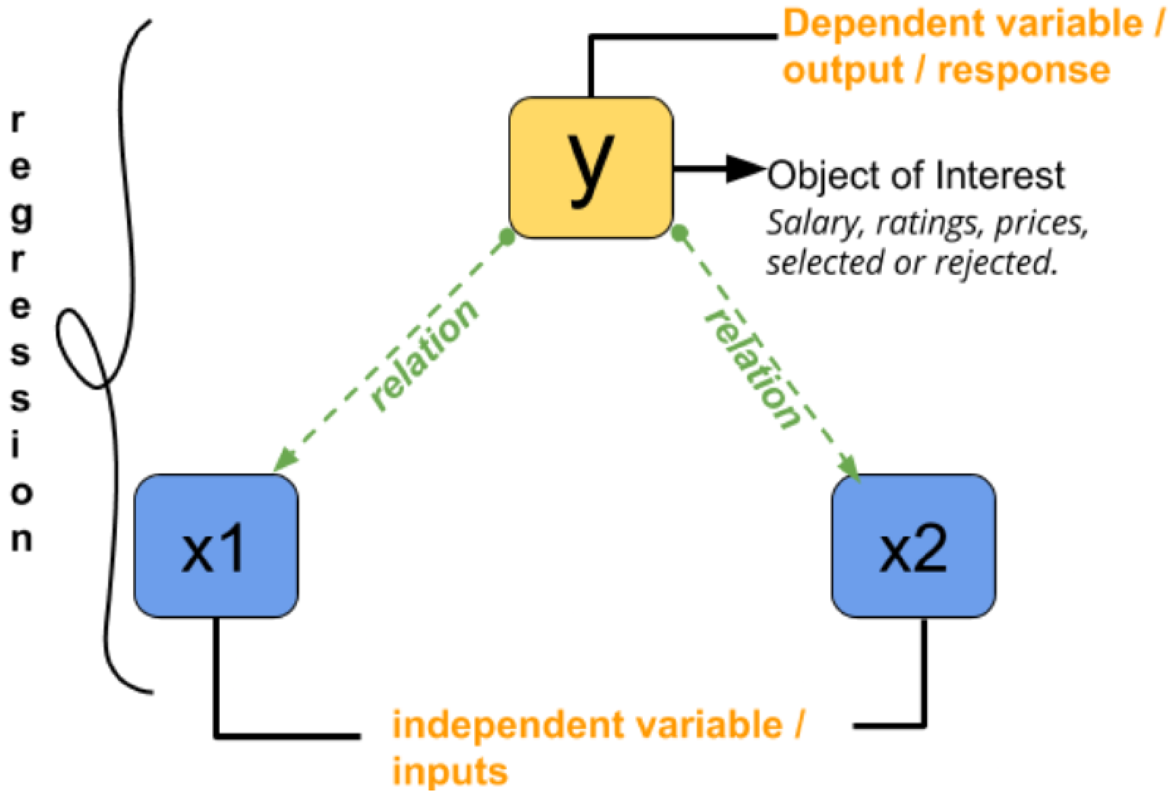
statistics.md

TILs of Statistics

1 June 2020

- Regression - formulates a relationship among variables of order >= 1.
- Check following image, if `y` is the object of interest, then regression is `finding a function that maps the variables to the object of interest (y) sufficiently well.`

```
# y as a function of x
y = f(x1, x2, x3, ..., xr)
```


Dependent variable / output / response
y
Object of Interest
Salary, ratings, prices, selected or rejected.
relation
relation
regression
x1
x2
independent variable / inputs

- Linear Regression - assuming a relation of order =1, between (x1,x2,x3 ..) and y.

```
y(i) = b0 + b1x1 + b2x2 + ... + brxr
```

where `f(x) = y(i)` is estimated regression function.

`b0, b1, b2 ...` = predicted weights, which captures the **dependencies** between the **input** and **output**.

1 June 2020

- For all observations in rows, `y(i)` should be as close to `y` as possible.
- `y - y(i)` is called **residual**.

- **REGRESSION IS ABOUT DETERMINING THE BEST PREDICTED WEIGHTS `(b0, b1 , b2 ... , br)` SUCH THAT THE `RESIDUAL` IS MINIMUM.** This is called **method of ordinary least sqaure.**
- **SUM OF SQAURED RESIDUAL (SSR)** = `SUM((y - y(i))^2)` , and best value for coef(b0, b1, b2, ...) are given by minimising the SSR.

3 June 2020

- The **coefficient of determination** (R^2) tells how much dependence of `y` is on `y(i)` . **HIGHER R^2 INDICATE BETTER FIT AND BETTER MODEL.**
- R^2 = 1 will give SSR = 0, i.e. perfect fit model.
- Simple linear regression of 1 variable means y is dpendeny on single variable x, therefore

```
# Linear function, b0 = intercept, b1 = slope of line
y = b0 + b1x
```

- Multiple linear regression is linear regression with two or more independent variables `(x)`.

```
y = b0 + b1x1 + b2x2 + ...
```

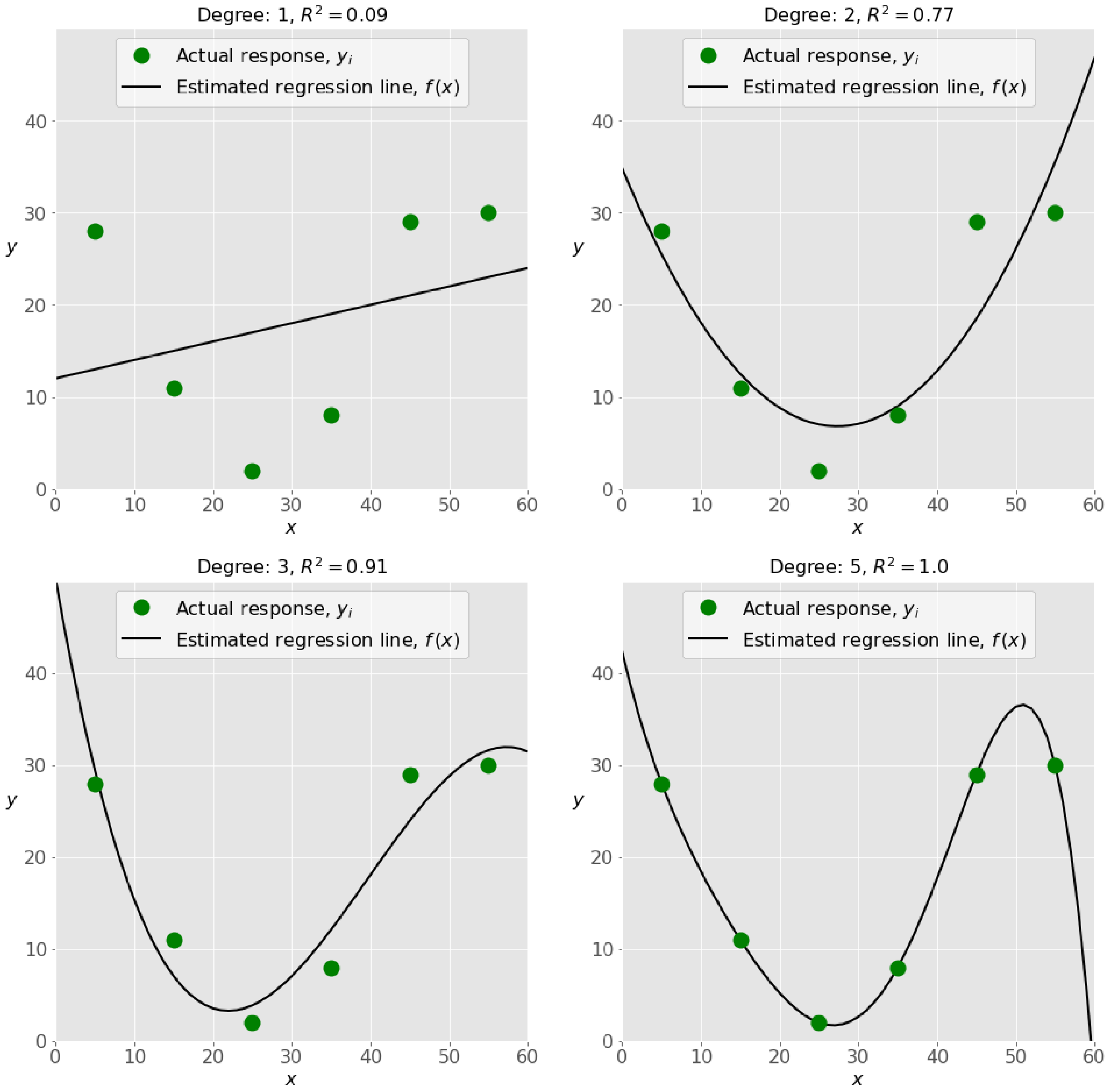
  - if `x` are plotted in different axes, the regression function becomes a plane, instead of a line. *(High level)*
- Polynomial regression is assuming polynomial relation between `y` and `x`.

```
y = b0 + b1x + b2(x^2) + ...
```

  - Polynomial equation can be converted to linear equation by, `x^2 = x2` , `x^3 = x3, ...`
- When implementing regression we want to get as close to `y` as possible such that `SSR is 0`.

4 June 2020

- You have the choice to choose the degree to the fit the model and to adjust its accuracy or tunings.
- The following figure depicts the terms **underfitting** and **overfitting** perfectly.



Graph	R^2 value	Degree	Description
top-left	0.009	1	model not perfectly represented
top-right	0.77	2	nice fit, and can be extrapolated to unknown/future inputs
bottom-left	0.91	3	accurate fit, but susceptible to errors when used for unknown inputs i.e. cannot be generalised
bottom-right	0.99	5	perfect fit, but cannot be used for new inputs. **cannot be generealised**

5 June 2020

- Simple linear regression with `sklearn` package in python, comprises of:
    i. import packages & classes

```
# import LinearRegression from sklearn
from sklearn import LinearRegression
```

ii. provide data to work with.
   - The inputs (x, regressors, independent var, actual values) should be 2-dimensional array `(1 column and multiple rows)`, use `.reshape(-1,1)` to convert.

iii. create a model and fit it.

```
# create a variable model of LinearRegression class
model = LinearRegression()
mode.fit(x,y)
```

iv. get results

```
# (R^2) Coefficient of determination
r_sq = model.score(x,y)
```

v. predict responses to unkown values.

```
# use .predict() to predict response for x values.
# y_predict = predictions of value y
y_predict = model.predict(x) # where x is array of new values/inputs.
```

- Check out this article for more in-depths - Link