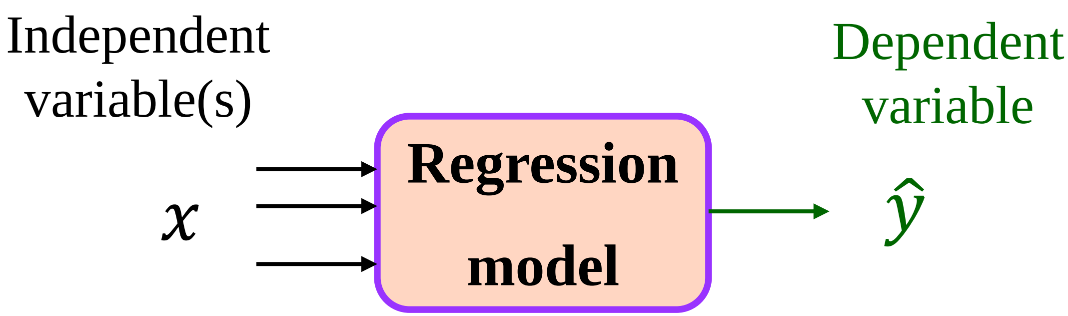
Regression

- Simple Linear Regression
- Multiple Linear Regression
- Polynomial “Linear” Regression

Regression analysis

- In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships among variables.

- It includes many techniques for modeling and analyzing several variables, when the focus is on the **relationship between a dependent variable and one or more independent variables** (or 'predictors’).



More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any of the independent variable is varied, while the other independent variables are held fixed.

[Regression analysis, Wikipedia, https://en.wikipedia.org/wiki/Regression_analysis]

Regression

- The regression is an approach **to model** the relationship between a scalar response (**dependent variable / regressor**) and one or more input variables (**independent variables**).
- Regression models (both linear and non-linear) are **machine-learning models;** used for **predicting/forecasting.**
- Regression models are used for **predicting a real value** (salary, stock prices, customer lifetime, sales, house prices). If the independent variable is time, then you are forecasting future values. Otherwise, the model is predicting present but unknown values.

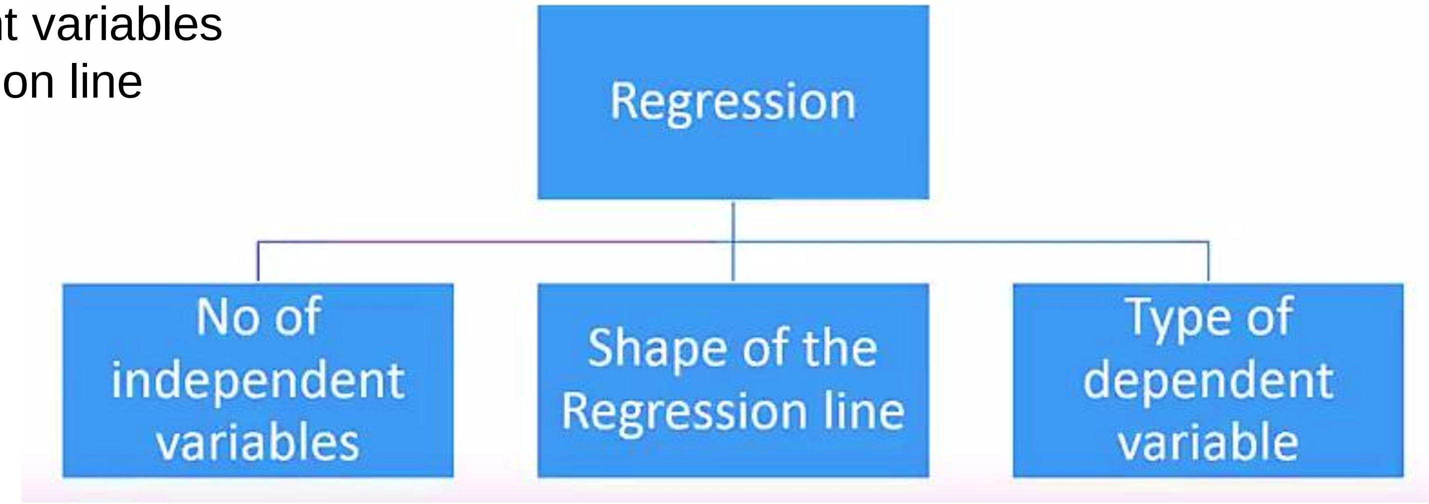
- A regression model must learn the correlation between data.

The case of **one input** variable (explanatory variable; independent variable) is called **simple linear regression**.

For **more input** variables (explanatory variables; independent variable), the process is called **multiple linear regression**.

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics:

- number of independent variables
- type of dependent variables
- shape of regression line



[7 Types of Regression Techniques you should know!, https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/]

Simple Linear Regression

$$\hat{y} = ax + b$$

y – the output / dependent variable (DV)

x – the input / independent variable (IV)

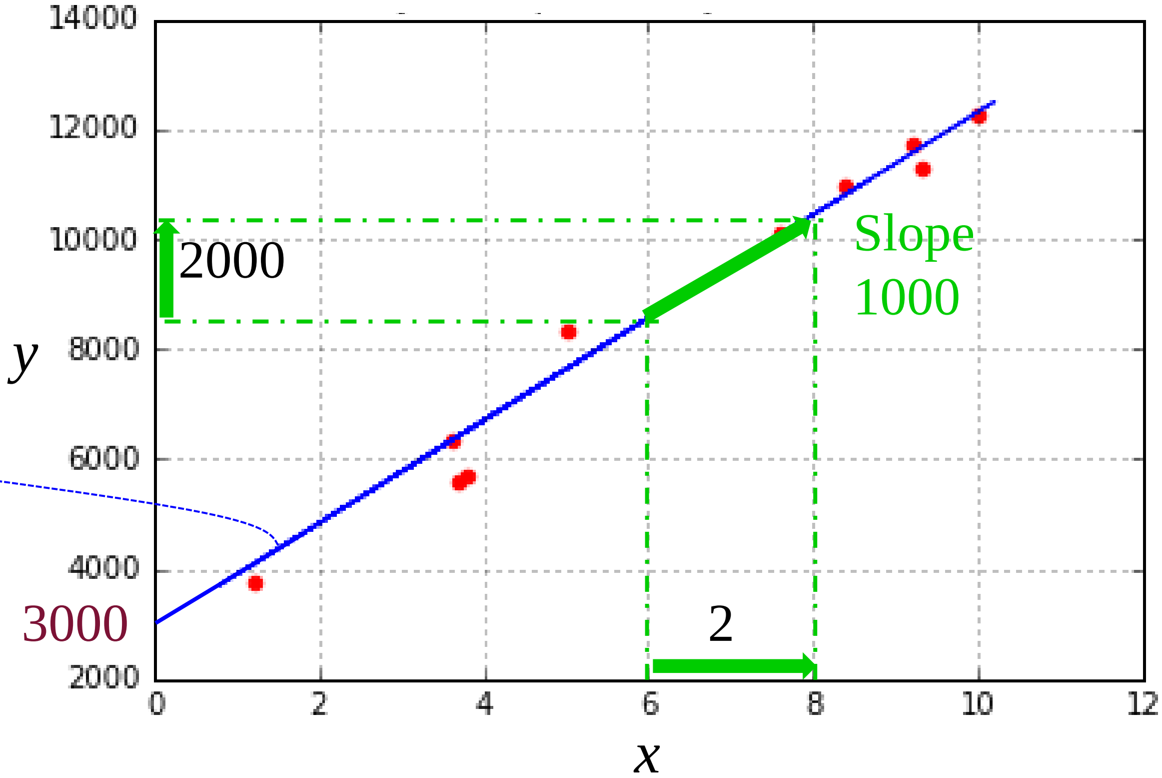
a – coefficient (**slope**)

b - constant (**intercept**)

intercept – y value where the line cuts the Y axis

Regression line

$$\hat{y} = 1000x + 3000$$



Red dots – facts; blue line – best fits the facts (the data) – linear regression

Linear regression: a trend line that best fits the data

Case study – build a simple regression model to

predict the salary in a company for a new employee

according with years of experience in the workforce.

The model will be built based on a set of data

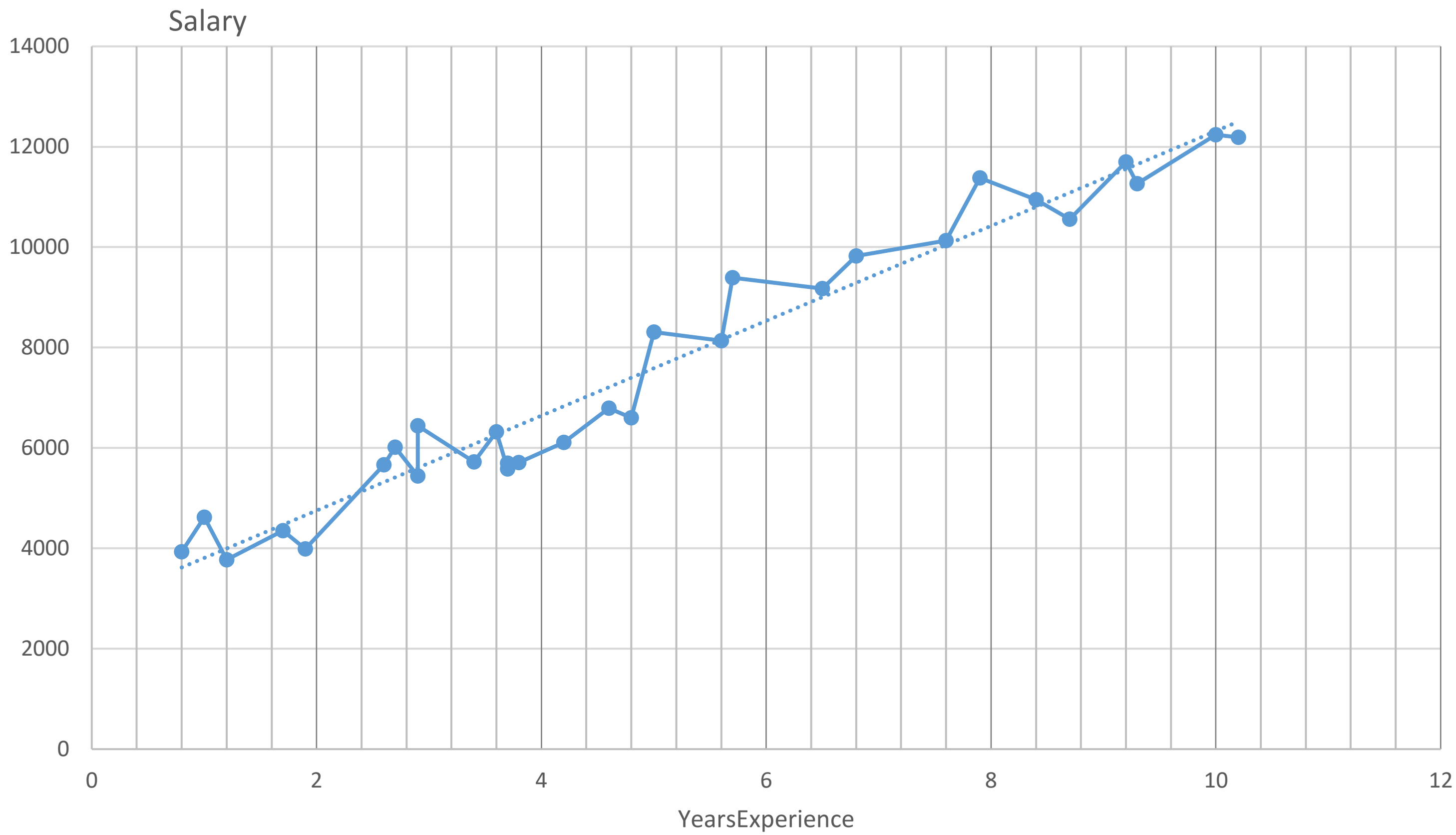
- 30 observations from that company

Data set

What is the correlation between years of experience and the salary?

YearsExperience	Salary
0.8	3934.3
1	4620.5
1.2	3773.1
1.7	4352.5
1.9	3989.1
2.6	5664.2
2.7	6015
2.9	5444.5
2.9	6444.5
3.4	5718.9
3.6	6321.8
3.7	5579.4
3.7	5695.7
3.8	5708.1
4.2	6111.1
4.6	6793.8
4.8	6602.9
5	8308.8
5.6	8136.3
5.7	9394
6.5	9173.8
6.8	9827.3
7.6	10130.2
7.9	11381.2
8.4	10943.1
8.7	10558.2
9.2	11696.9
9.3	11263.5
10	12239.1
10.2	12187.2

Data set representation



YearsExperience	Salary
0.8	3934.3
1	4620.5
1.2	3773.1
1.7	4352.5
1.9	3989.1
2.6	5664.2
2.7	6015
2.9	5444.5
2.9	6444.5
3.4	5718.9
3.6	6321.8
3.7	5579.4
3.7	5695.7
3.8	5708.1
4.2	6111.1
4.6	6793.8
4.8	6602.9
5	8308.8
5.6	8136.3
5.7	9394
6.5	9173.8
6.8	9827.3
7.6	10130.2
7.9	11381.2
8.4	10943.1
8.7	10558.2
9.2	11696.9
9.3	11263.5
10	12239.1
10.2	12187.2

Verify on the dataset

$$\hat{y} = 1000x + 3000$$

Data point $x = 5, \quad y = 8308.8$

Predicted value

$$\hat{y} = 1000 \cdot 5 + 3000 = 8000$$

Error

$$\hat{y} - y = 8308.8 - 8000 = 308.8$$

Using the regression model (for new inputs)

Determine the Salary (dependent variable) for a new value of Years Experience (independent variable)

$$x = 9$$

$$\hat{y} = 1000 \cdot 9 + 3000 = 12000$$

YearsExperience	Salary
0.8	3934.3
1	4620.5
1.2	3773.1
1.7	4352.5
1.9	3989.1
2.6	5664.2
2.7	6015
2.9	5444.5
2.9	6444.5
3.4	5718.9
3.6	6321.8
3.7	5579.4
3.7	5695.7
3.8	5708.1
4.2	6111.1
4.6	6793.8
4.8	6602.9
5	8308.8
5.6	8136.3
5.7	9394
6.5	9173.8
6.8	9827.3
7.6	10130.2
7.9	11381.2
8.4	10943.1
8.7	10558.2
9.2	11696.9
9.3	11263.5
10	12239.1
10.2	12187.2

Quality of regression

Errors and residuals

y - target (ground truth, original, observed) **$\hat{y}$ - predicted (estimated value)**

In statistics and optimization, **errors** and **residuals** are two closely related and easily confused measures of the deviation of an observed value of an element of a statistical sample from its "theoretical value". [https://en.wikipedia.org/wiki/Errors_and_residuals]

The **residual** of an observed value is the difference between the observed value and the *estimated* value of the quantity of interest. Residuals are the difference between any data point and the regression line

In a linear regression context, ***residuals applies to the dataset*** *(training, test, validation):*

$y - \hat{y}$. ***The residuals are observable.***

The **error** (or **disturbance**) of an observed value is the deviation of the observed value from the (unobservable) *true* value of a quantity of interest.

In a linear regression context, ***error refers to the results in the model utilization phase****.* *(true value – predicted value)* y_{true} - $\hat{y}$*. Because we really don't know the true value****, the error is unknown.***

In the context of **machine learning**, the term **"error"** (singular) means the difference between predicted and target values,

$$error = \hat{y} - y$$

and the term "residual(s)" is practically almost never used.

Errors

For a dataset with m examples: $y^{(i)}$ denotes the i^{th} example (the target)

$\hat{y}^{(i)}$ (the prediction)

Error $\hat{y}^{(i)} - y^{(i)}$

Squared error $\left(\hat{y}^{(i)} - y^{(i)}\right)^2$

Mean squared-error $$\frac{1}{m}\sum_{i=1}^{m}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

Root mean squared-error $$\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2}$$

Ordinary Least Square

We want to build a simple linear regression model

$$\hat{y} = ax + b$$

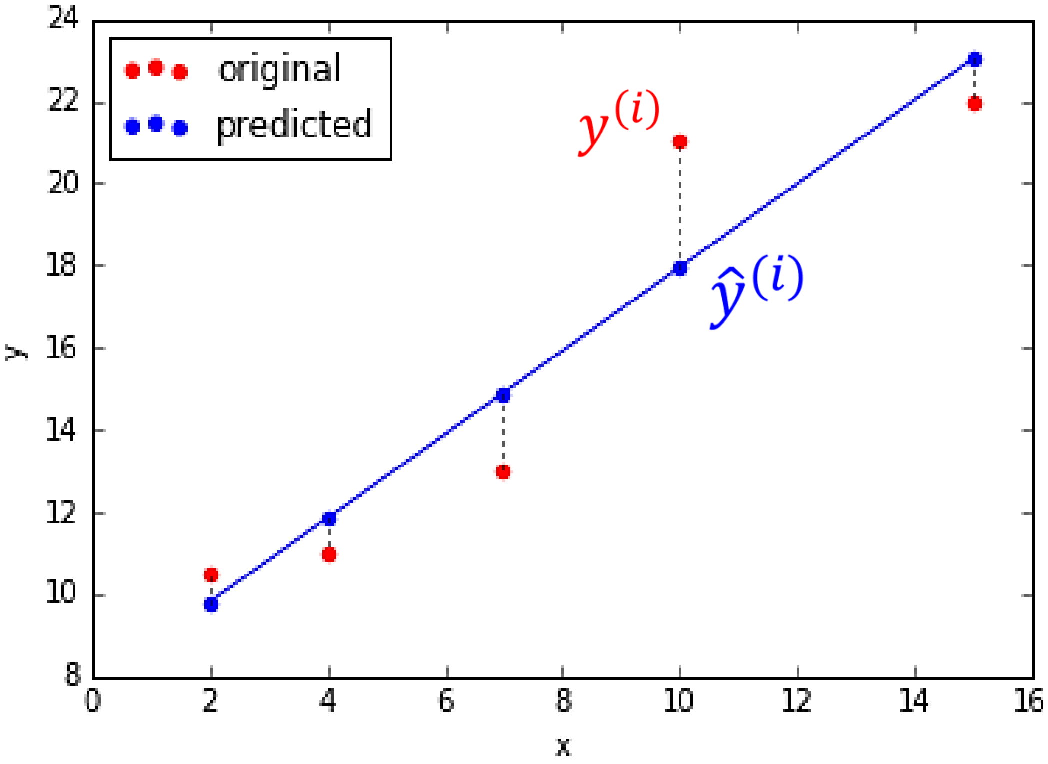
How can the model parameters be estimated (calculated)?

$$a = ? \quad b = ?$$

Ordinary Least Squares (OLS) is a method used to estimate the parameters (coefficients) of a linear regression model.

The goal of OLS is to **find the best-fitting line through the data points** by **minimizing the sum of the squared errors** between the values predicted by the linear model and the target values

Ordinary Least Square



Sum of the squares error

$$\sum_{i=1}^{n} \left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

to be minimized

The sum of the *squares* of the errors is used instead of the absolute values of the error because this allows the residuals to be treated as a continuous differentiable quantity.

Outlying points can have a disproportionate effect on the fit, a property which may or may not be desirable depending on the problem at hand.

Coefficient of determination R^2

The coefficient of determination, commonly denoted as R^2 (R squared) is a statistical measure used to assess how well a linear regression model explains the variability in the dependent variable.

In simpler terms, R^2 tells us how much of the variation in the outcome (dependent variable) is explained by the predictor variables (independent variables) in the model.

R^2 values range from 0 to 1:

- **R^2 = 1**: The model perfectly explains all the variation in the dependent variable.
- **R^2 =0**: The model explains **none** of the variation (i.e., the model's predictions are no better than the mean of the data).
- **Closer to 1**: The model explains a large proportion of the variation in the dependent variable.
- **Closer to 0**: The model explains very little of the variation.

Coefficient of determination R^2

Interpretation:

- R^2 represents the **percentage** of the total variation in the dependent variable that can be explained by the independent variables in the model.

- For example, if $R^2 = 0.85$, this means that 85% of the variation in the outcome is explained by the predictors, while the remaining 15% is due to factors not included in the model (or noise).

$$R^2 \equiv 1 - \frac{SS_{res}}{SS_{tot}} = \frac{SS_{tot} - SS_{res}}{SS_{tot}}$$

$$SS_{res} = \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

Residual sum of squares

$$SS_{tot} = \sum_{i=1}^{m} (y_i - \bar{y})^2$$

Total sum of squares

$$\bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$$ Mean of the data

Python code for linear regression

```python
# %% Simple linear regression: 30 observations: years of experience - salar

# Importing the library
import numpy as np    # math tools
import matplotlib.pyplot as plt  # for plotting charts
import pandas as pd  # import and manage datasets

# %% Importing the dataset  (.csv)
dataset = pd.read_csv('Salary_Data_ICSDC.csv')
# print('\n**  The data set is: \n\n', dataset)
#   x = dataset.iloc[:,0].values  # IV (independent variable)
x = dataset.iloc[:,:-1].values  # both are correct for x
y = dataset.iloc[:,1].values  # DV (dependent variable)

# %% Split the dataset into the Training set and Test set
from sklearn.cross_validation import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 1/3,
    random_state=0)

# %% Fitting the Simple Linear Regression to the Training set
from sklearn.linear_model import LinearRegression
# from .linear_model library import LinearRegresion class

regressor = LinearRegression()  # build our own object named "regressor"
regressor.fit(x_train, y_train) # use the method .fit of our object

# %% Predicting for the x_test
y_pred = regressor.predict(x_test)
```

```
# %% Plot some graph
plt.close('all')  # close already existing plots
# Trainig set
plt.figure()
plt.scatter(x_train, y_train, color = 'green')
plt.plot(x_train, regressor.predict(x_train),color = 'blue')
plt.title('Salary vs. Experience(years) - Training set')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
# Test set
plt.figure()
plt.scatter(x_test, y_test, color = 'red')
plt.plot(x_train, regressor.predict(x_train),color = 'blue')
plt.title('Salary vs. Experience(years) - Test set')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
# All data
plt.figure()
plt.scatter(x_train, y_train, color = 'green')
plt.scatter(x_test, y_test, color = 'red')
plt.plot(x_train, regressor.predict(x_train),color = 'blue')
plt.title('Salary vs. Experience(years) - All data')
plt.xlabel('Years of Experience')
plt.ylabel('Salary')
plt.show()
```

Results

Training set

Slope = 934.59424431

Intercept = 2961.9974976967906

Salary = 935 * Years + 2962

MSE = 368529.5

RMSE = 607.1

$R^2 = 0.9382$

Results

Test set

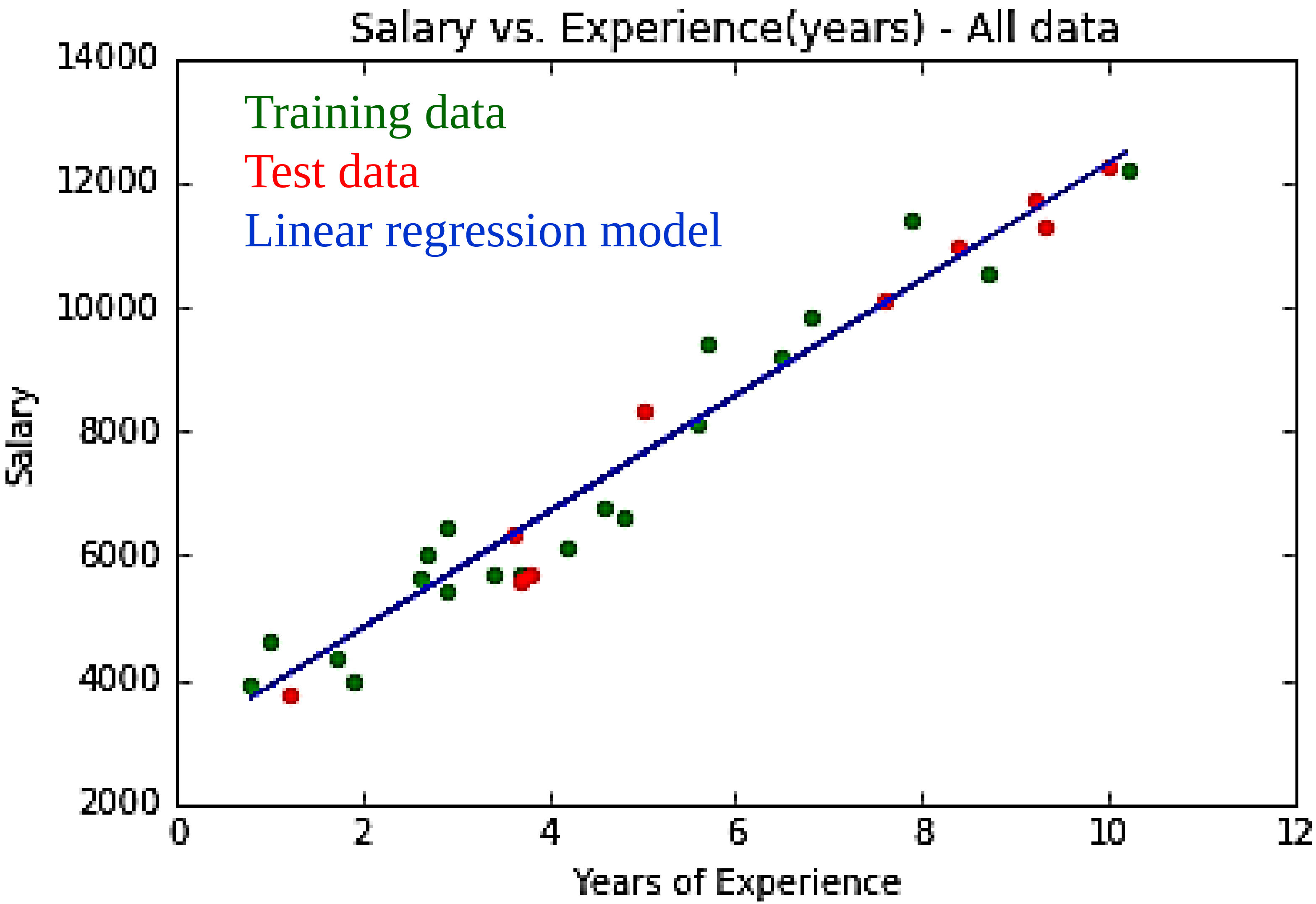
MSE = 210260.4
RMSE = 458.5

$R^2 = 0.9749$

test	[3773.1, 12239.1, 5708.1, 6321.8, 11696.9, 10943.1,11263.5, 5579.4, 8308.8, 10130.2]
pred	[4083.5, 12307.9, 6513.5, 6326.5, 11560.3, 10812.6,11653.7, 6420.0, 7635.0, 10064.9]
pred-test	[310.4, 68.8, 805.4, 4.7, -136.6, -130.5, 390.2, 840.6, -673.8, -65.2]

Results

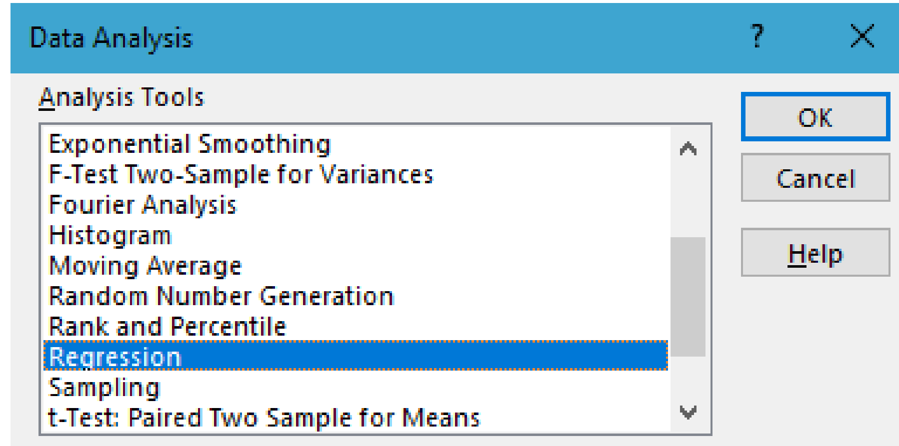
All data

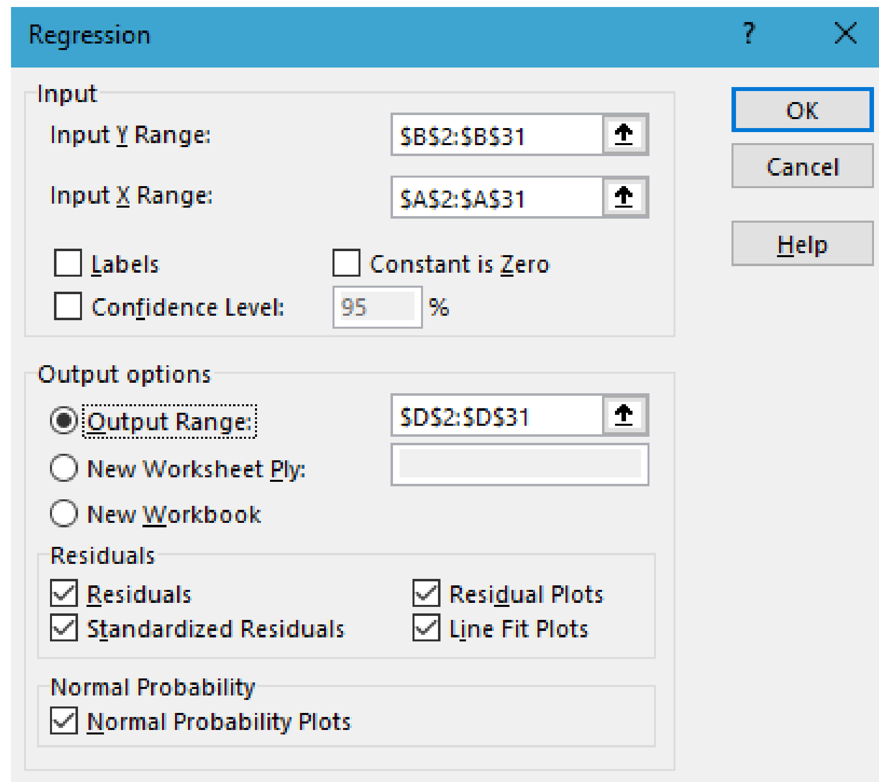


Regression in Excel

	A	B
1	YearsExperienc	Salary
2	0.8	3934.3
3	1	4620.5
4	1.2	3773.1
5	1.7	4352.5
6	1.9	3989.1
7	2.6	5664.2
8	2.7	6015
9	2.9	5444.5
10	2.9	6444.5
11	3.4	5718.9
12	3.6	6321.8
13	3.7	5579.4
14	3.7	5695.7
15	3.8	5708.1
16	4.2	6111.1
17	4.6	6793.8
18	4.8	6602.9
19	5	8308.8
20	5.6	8136.3
21	5.7	9394
22	6.5	9173.8
23	6.8	9827.3
24	7.6	10130.2
25	7.9	11381.2
26	8.4	10943.1
27	8.7	10558.2
28	9.2	11696.9
29	9.3	11263.5
30	10	12239.1
31	10.2	12187.2

Data
Data Analysis





Regression in Excel - Results

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	Standard Residuals
1	3618.715875	315.5841248	0.554859429
2	3807.715122	812.7848783	1.42903688
3	3996.714368	-223.6143681	-0.393158371
4	4469.212484	-116.7124842	-0.205203675
5	4658.211731	-669.1117306	-1.176431016
6	5319.709093	344.4909069	0.60568328
7	5414.208716	600.7912837	1.056310132
8	5603.207963	-158.7079627	-0.279040049
9	5603.207963	841.2920373	1.479158115
10	6075.706079	-356.8060788	-0.627335792
11	6264.705325	57.09467477	0.100383752
12	6359.204948	-779.8049484	-1.371051628
13	6359.204948	-663.5049484	-1.166573182
14	6453.704572	-745.6045717	-1.310920589
15	6831.703065	-720.6030645	-1.266962985
16	7209.701557	-415.9015574	-0.731237354
17	7398.700804	-795.8008038	-1.399175512
18	7587.70005	721.0999498	1.267836607
19	8154.69779	-18.39778953	-0.03234696
20	8249.197413	1144.802587	2.012789806
21	9005.194398	168.6056015	0.296442059
22	9288.693268	538.6067319	0.946977367
23	10044.69025	85.50974618	0.150343079
24	10328.18912	1053.010877	1.851401789
25	10800.68724	142.4127605	0.250389854
26	11084.18611	-525.9861092	-0.924787811
27	11556.68423	140.2157748	0.246527118
28	11651.18385	-387.6838485	-0.68162503
29	12312.68121	-73.58121097	-0.12937035
30	12501.68046	-314.4804574	-0.552918963

SUMMARY OUTPUT

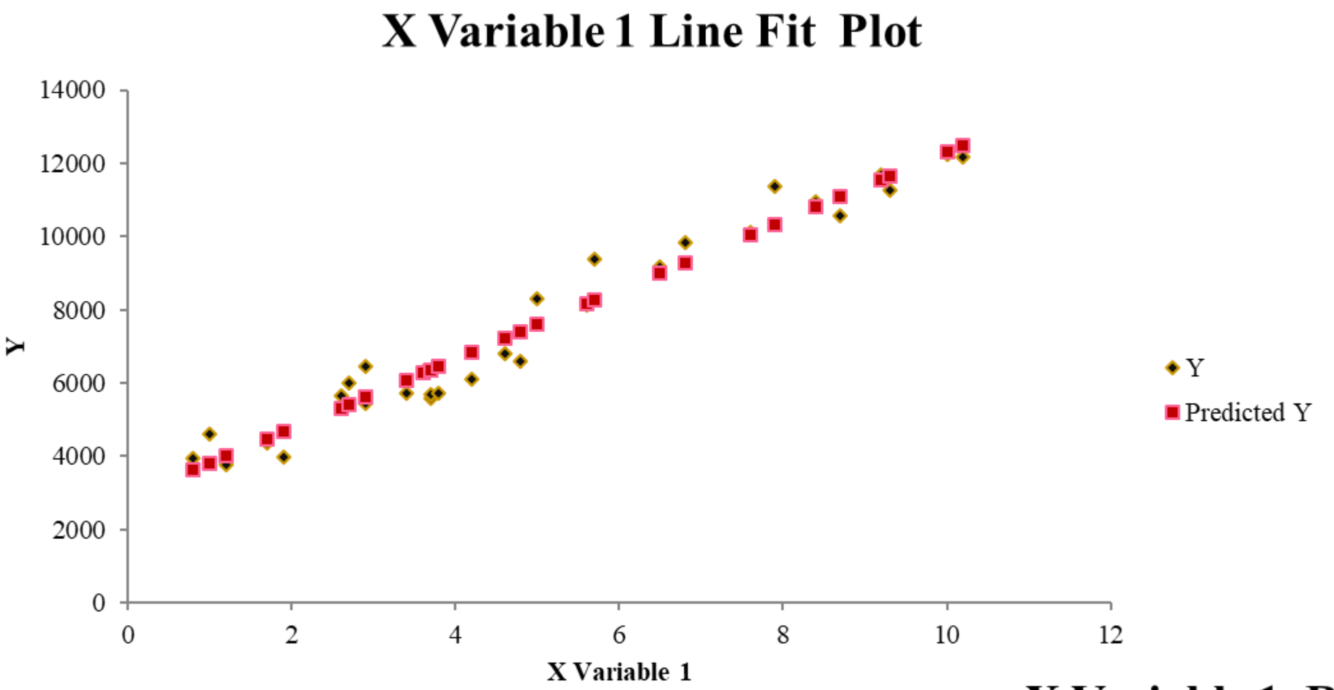
Regression Statistics	
Multiple R	0.978242
R Square	0.956957
Adjusted R Square	0.955419
Standard Error	578.8315
Observations	30

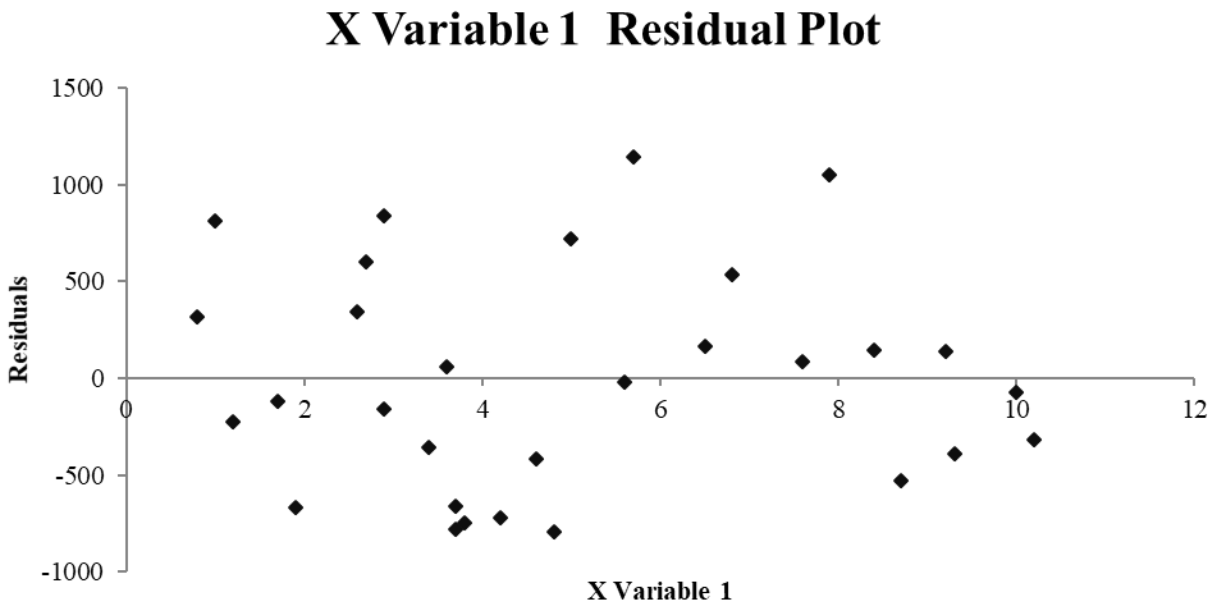
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	208568493	2.09E+08	622.5072	1.1431E-20
Residual	28	9381285.517	335045.9		
Total	29	217949778.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	2862.71889	217.3096462	13.17346	1.6E-13	2417.58026	3307.857521
X Variable 1	944.9962321	37.87545742	24.95009	1.14E-20	867.411875	1022.58059

Regression in Excel Results



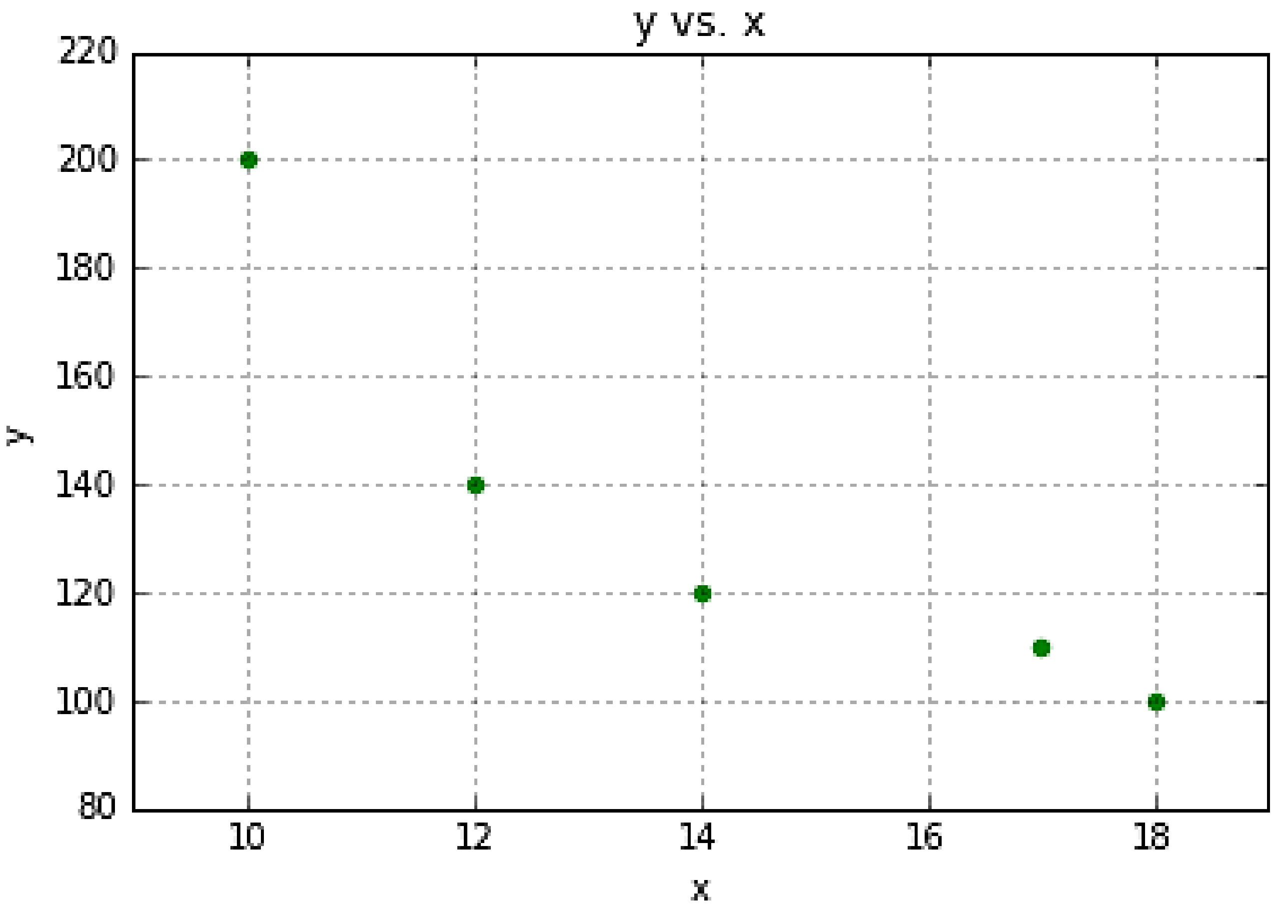


Exercise

Excel exercise

Using the OLS linear regression to model the $y(x)$ relation, the result is:

slope = -10.8;
intercept = 287.



a) What is the equation of the linear regression model?
b) Plot the linear model on the same diagram.
c) What are the errors?
d) What is the interpretation of the following error measures:

MSE = 218.2; RMSE = 14.77; $R^2 = 0.8274$

Multiple Linear Regression

$$\hat{y} = a_1 x_1 + a_2 x_2 + \dots a_n x_n + b$$

y – dependent variable (DV) / regressor

x_1, x_2, x_3, ... , x_n - independent variables (IVs) / predictors

a_1, a_2, a_3, ... , a_n - coefficients

b - constant

5 methods of building multiple linear regression models:

1. All–in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

(Methods 2–4: Stepwise regression)

[Kirill Ermenko, Building a Model (Step-By-Step), Data Science Training, https://www.google.ro/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=2ahUKEwi50pnIoZfeAhWRKCwKHdLKBIsQFjAAegQICxAC&url=https%3A%2F%2Fwww.superdatascience.com%2Fwp-content%2Fuploads%2F2017%2F02%2FStep-by-step-Blueprints-For-Building-Models.pdf&usg=AOvVaw0C8l04IYGkS6i23PeeLrqg]

Multiple Linear Regression - *Backward elimination*

Usually, we are using all dependent variables; but is this the optimal model?

Some independent variables (IV) can be highly statistically significant with great impact (effect) on the DV (dependent variable)

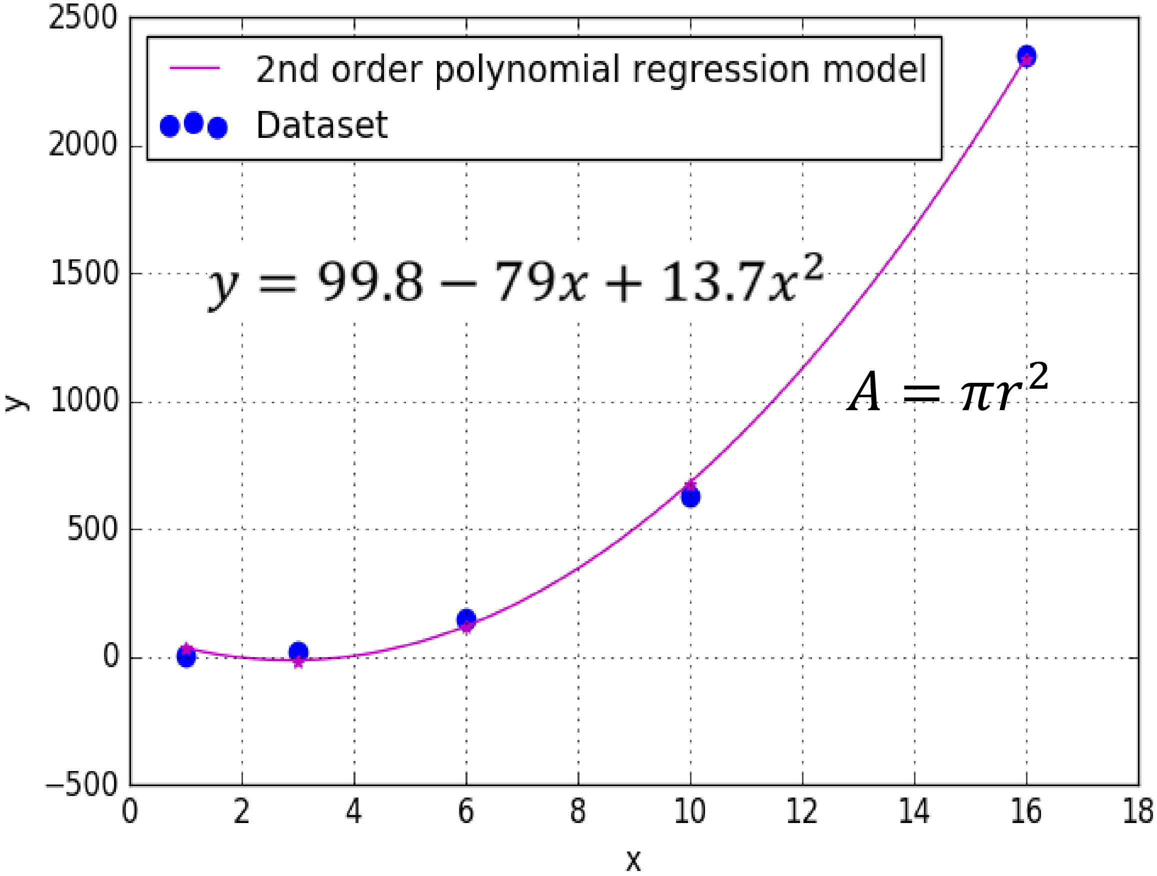
Some IVs are not statistically significant at all – should be removed from the model.

Find a team of optimal IVs, where each IV of the team has great impact on the DV (statistically significant)

Polinomial „linear” regression

One independent variable x

$$\hat{y} = a_1 x + a_2 x^2 + \ldots a_n x^n + b$$



Can be seen as a special case of a multiple linear regression – from the point of view of a_i coefficient

independent variables

$x, x^2, \cdots, x^n$

In statistics, **polynomial regression** is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an m^{th} degree polynomial in x.

Polynomial regression fits a **nonlinear relationship** between the value of x and the corresponding conditional mean of y, denoted $E(y \mid x)$.

Although polynomial regression fits a nonlinear model to the data,
as a statistical estimation problem it is linear,
in the sense that the regression function $E(y \mid x)$ is linear in the unknown parameters (a_o, a_1, ..., a_m) that are estimated from the data.

For this reason, polynomial regression is considered to be a special case of multiple linear regression *[https://en.wikipedia.org/wiki/Polynomial_regression]*

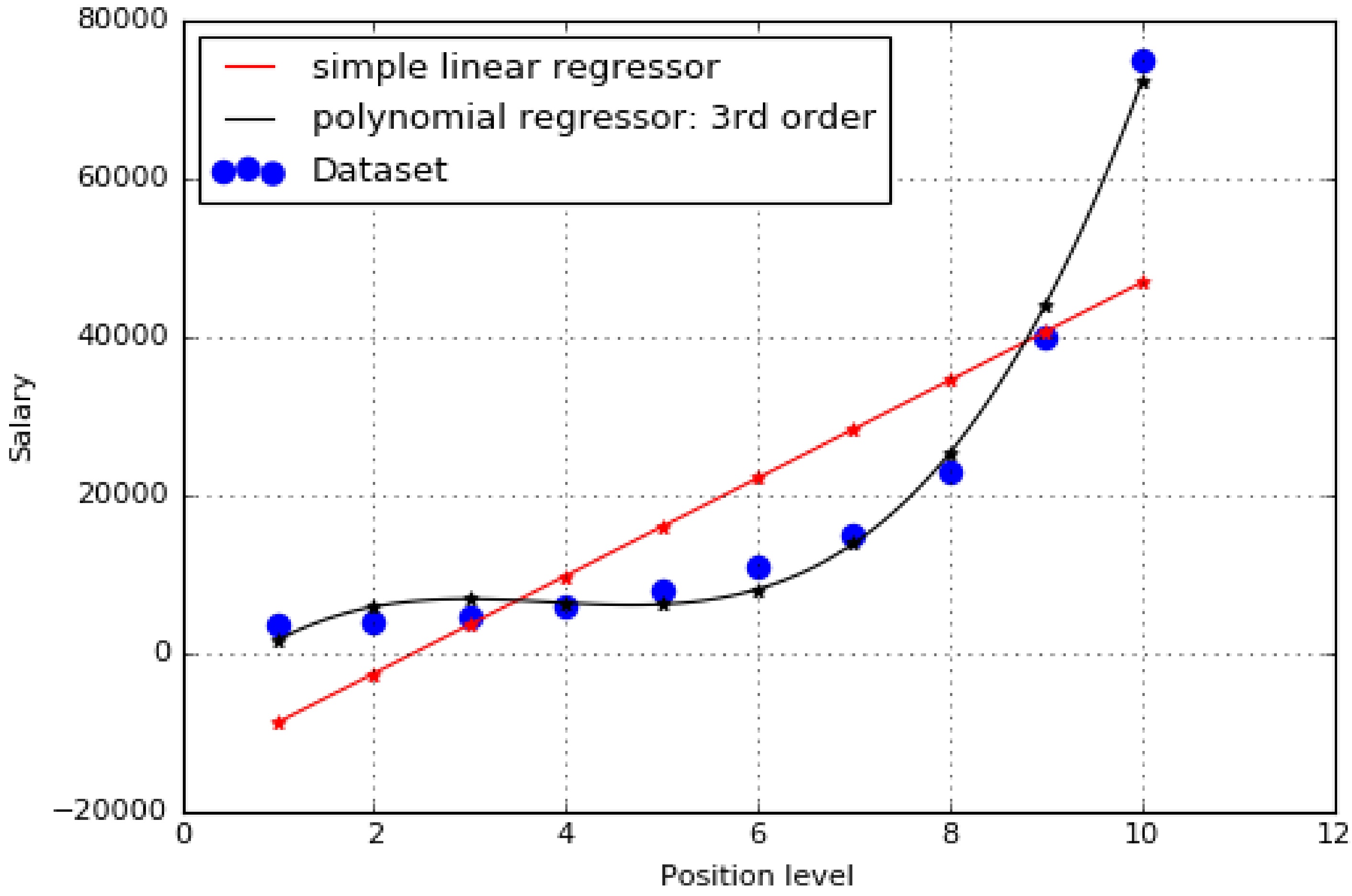
A linear combination from the coefficient point a view.
In fact, **the problem is to determine the coefficients**.

Data set: Position- salary

Position	Level	Salary
Business Analyst	1	3500
Junior Consultant	2	3900
Senior Consultant	3	4500
Manager	4	5800
Country Manager	5	8000
Region Manager	6	11000
Partner	7	15000
Senior Partner	8	23000
C-level	9	40000
CEO	10	75000

What is the best polynomial model Position (Level) – Salary?

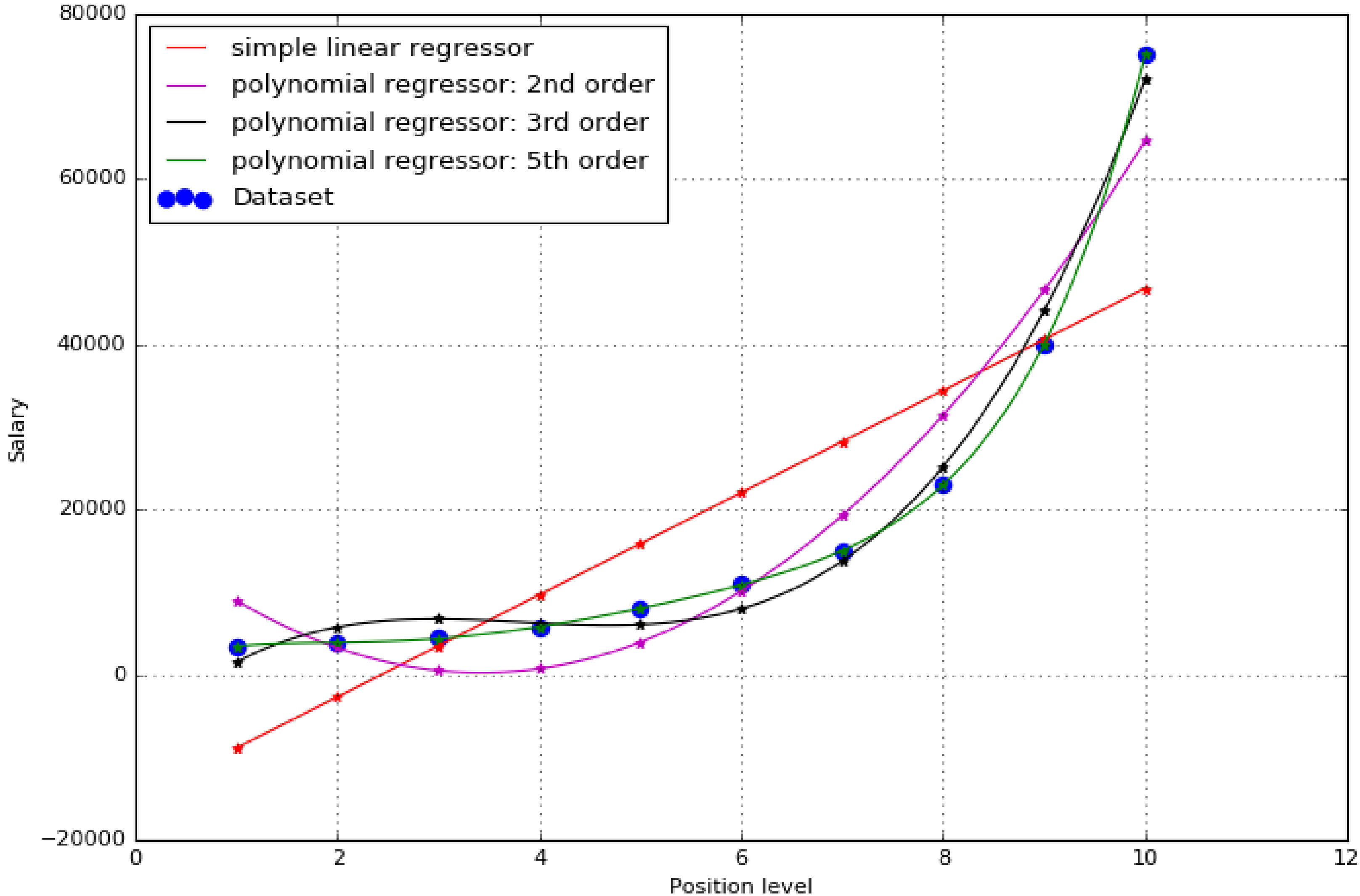
1^{st} order vs 3^{rd} order polynomial model



1^{st} : Salary = - 15006 + 6178 Level

3^{rd} : Salary = - 7747 + 12408 Level – 3412 $Level^2$ + 297 $Level^3$

Comparison – different polynomial model



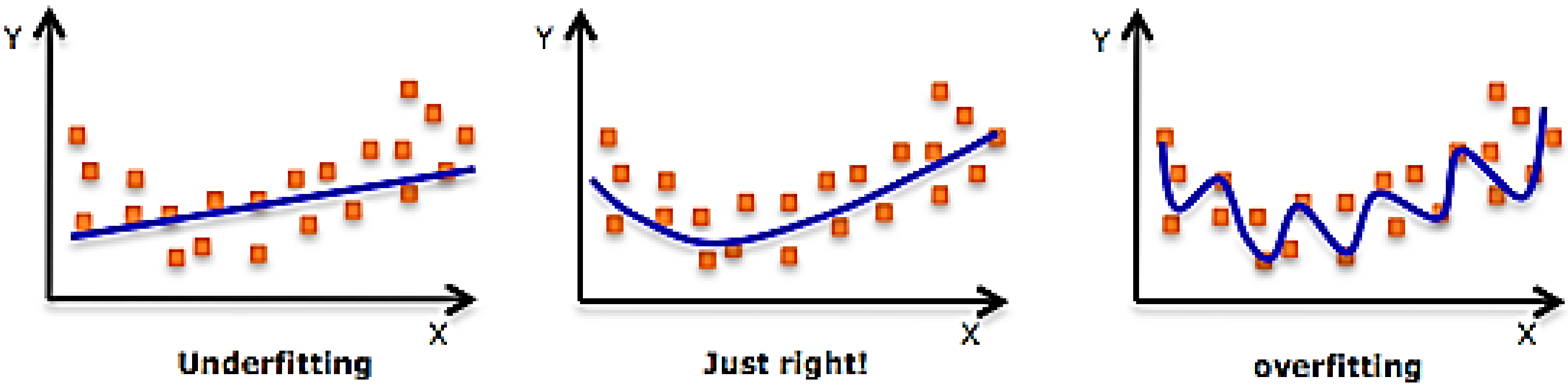
Order	1st (linear)	2nd	3rd	5th
R^2	0.6775504	0.9292263	0.9878567	0.9999947

Important aspects

- While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in **over-fitting**.

- Always plot the relationships to see the fit and focus on making sure that the **curve fits the nature of the problem**

- Look out for curve towards the **ends** and see whether those shapes and trends make sense. Higher polynomials can end up producing weird results on extrapolation.



[7 Types of Regression Techniques you should know!, https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/]

Exercise

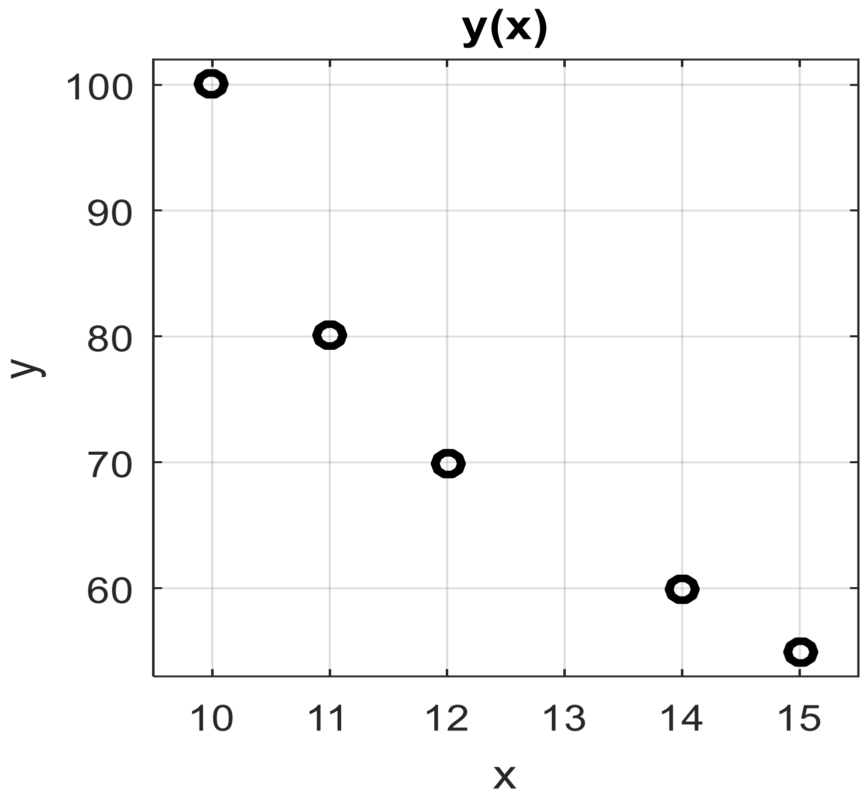
For the dataset presented in the next plot, a linear regression model $\widehat{y_l} = ax + b,\ b = 175,\ a = -8$, and a quadratic regression model $\widehat{y_q} = a_1 x + a_2 x^2 + b$ $b = 470;\ a_1 = -56; a_2 = 1.9$, were developed.

a) **0.5p** Which is the equation of the linear model? Plot the regression line.

b) **0.5p** Which is the equation of the quadratic model? Plot the regression curve.

c) **1p** Which is the predicted value of the dependent variable *y* for the independent variable *x* = 12 for both models. Plot this data points.

d) **1p** Which of the two models is more accurate? Why?



Excel exercise