REGRESI LINEAR: TEORI DAN PRAKTIK

Isram Rasal, S.T., MMSI., M.Sc

Machine Learning Course

16/02/2019

Laboratorium Deep Learning

Pendahuluan

- Penggunaan statistika dalam mengolah data penelitian berpengaruh terhadap tingkat analisis hasil penelitian.
- Penelitian-penelitian dalam bidang ilmu pengetahuan alam (science) yang menggunakan perhitungan-perhitungan statistika, akan menghasilkan data yang mendekati benar jika memperhatikan tata cara analisis data yang digunakan.
- Dalam memprediksi dan mengukur nilai dari pengaruh satu variabel (bebas/independent/predictor) terhadap variabel lain (tak bebas/dependent/response) dapat digunakan *uji regresi*.

Pendahuluan (2)

- Analisis/uji regresi merupakan suatu kajian dari hubungan antara satu variabel, yaitu variabel yang diterangkan (***the explained variabel***) dengan satu atau lebih variabel, yaitu variabel yang menerangkan (***the explanatory***).
- Apabila variabel bebasnya hanya satu, maka analisis regresinya disebut dengan **regresi linear sederhana**.
- Apabila variabel bebasnya lebih dari satu, maka analisis regresinya dikenal dengan **regresi linear berganda**. Dikatakan berganda karena terdapat beberapa variabel bebas yang mempengaruhi variabel tak bebas

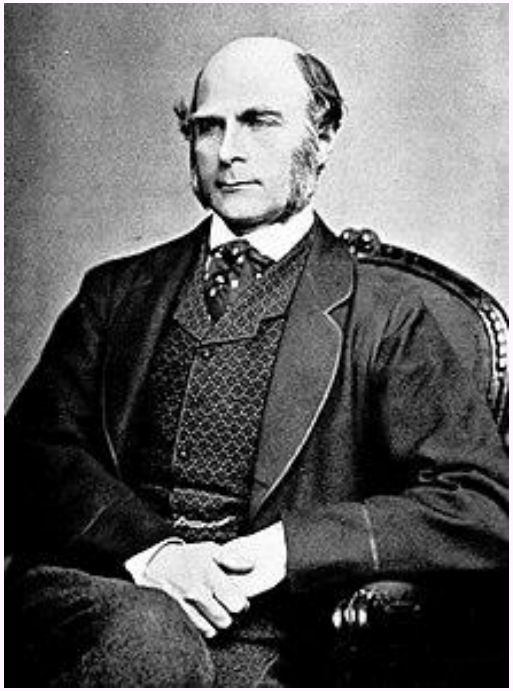
Pendahuluan (3)

- Analisis/uji regresi banyak digunakan dalam perhitungan hasil akhir untuk penulisan karya ilmiah/penelitian.
- Hasil perhitungan analisis/uji regresi akan dimuat dalam kesimpulan penelitian dan akan menentukan apakah penelitian yang sedang dilakukan berhasil atau tidak.
- Analisis perhitungan pada uji regresi menyangkut beberapa perhitungan statistika seperti uji signifikansi (uji-t, uji-F), anova dan penentuan hipotesis.
- Hasil dari analisis/ uji regresi berupa suatu persamaan regresi. Persamaan regresi ini merupakan suatu fungsi prediksi variabel yang mempengaruhi variabel lain

Sejarah Regresi Linear

Regresi pertama kali diteliti secara mendalam oleh seorang ilmuwan kenamaan di abad ke-19, yang juga seorang ahli statistika, astronomer dan antroplogis.

Beliau menemukan ilmu regresi saat meneliti beberapa spesies tanaman dan hewan.



Sir Francis Galton

16 February 1822 – 17 January 1911

Regresi Linear Sederhana

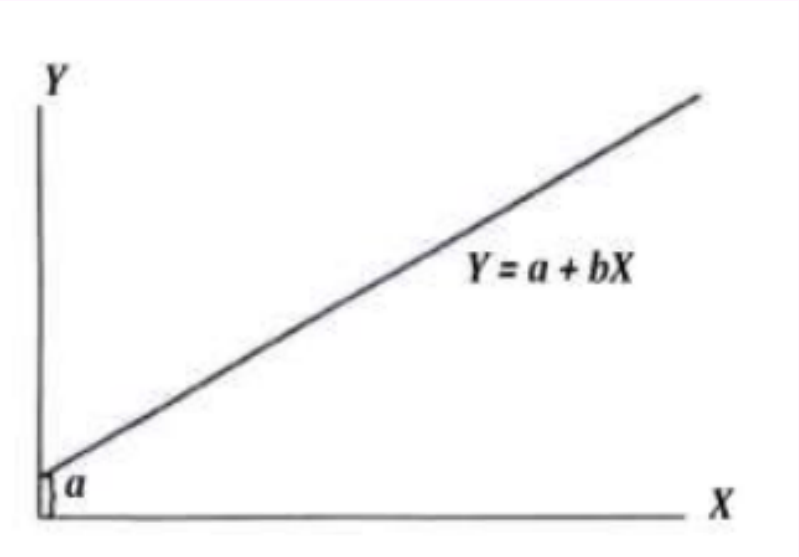
- Regresi Linear Sederhana adalah Metode Statistik yang berfungsi untuk menguji sejauh mana hubungan sebab akibat antara **Variabel Faktor Penyebab (X)** terhadap **Variabel Akibatnya**.
- Faktor Penyebab pada umumnya dilambangkan dengan **X** atau disebut juga dengan ***Predictor*** sedangkan Variabel Akibat dilambangkan dengan ***Y*** atau disebut juga dengan ***Response***.
- Regresi Linear Sederhana atau sering disingkat dengan SLR (*Simple Linear Regression*) juga merupakan salah satu Metode Statistik yang dipergunakan dalam produksi untuk melakukan peramalan ataupun prediksi tentang karakteristik kualitas maupun kuantitas.

Contoh Penggunaan

- Contoh penggunaan analisis Regresi Linear Sederhana dalam kegiatan produksi, antara lain :
  - Hubungan antara lamanya kerusakan mesin dengan kualitas produk yang dihasilkan
  - Hubungan jumlah pekerja dengan output yang diproduksi
  - Hubungan antara suhu ruangan dengan cacat produksi yang dihasilkan.

Model Persamaan Regresi Linear Sederhana

- Model Persamaan Regresi Linear Sederhana adalah seperti berikut ini :
  - Y = a + bX
- Dimana :
  Y = Variabel Response atau Variabel Akibat (Dependent)
  X = Variabel Predictor atau Variabel Faktor Penyebab (Independent)
  a = konstanta
  b = koefisien regresi (kemiringan/slope); besaran Response yang ditimbulkan oleh Predictor.



Model Persamaan Regresi Linear Sederhana (2)

- Besarnya konstanta a dan b dapat ditentukan menggunakan persamaan:

$$a = \frac{(\sum Y_i)(\sum X_i^2) - (\sum X_i)(\sum X_i Y_i)}{n\sum X_i^2 - (\sum X_i)^2}$$

$$b = \frac{n\ (\sum X_i\ Y_i) - (\sum X_i)\ (\sum Y_i)}{n\sum X_i^2\ - (\sum X_i)^2}$$

yang mana n = jumlah data

Langkah-langkah Analisis dan Uji Regresi Linier Sederhana

- Adapun langkah-langkah yang perlu dilakukan untuk melakukan analisis dan uji regresi linier sederhana adalah sebagai berikut:

1. Menentukan tujuan dari Analisis Regresi Linear Sederhana
2. Mengidentifikasi variabel *predictor* dan variabel *response*
3. Melakukan pengumpulan data dalam bentuk tabel
4. Menghitung X^2, Y^2, XY dan total dari masing-masingnya
5. Menghitung a dan b menggunakan rumus yang telah ditentukan
6. Membuat model Persamaan Garis Regresi
7. Melakukan prediksi terhadap variabel *predictor* atau *response*
8. Uji korelasi

Contoh Kasus

- "Terdapat suatu data penelitian tentang berat badan 10 mahasiswa yang diprediksi dipengaruhi oleh konsumsi jumlah kalori/hari".
- Bagaimana menganalisisnya?

1. Tujuan : "Apakah konsumsi jumlah kalori/hari mempengaruhi berat badan mahasiswa?"
2. X (variable bebas/predictor) = jumlah kalori/hari

   Y (variable tak bebas/response) = berat badan

Tabel Data

3. Melakukan pengumpulan data dalam bentuk tabel

Nama Mahasiswa	Kalori/ hari (X)	Berat Badan (Y)
Adi	530	89
Rudi	300	48
Didi	358	56
Budi	510	72
Intan	302	54
Putu	300	42
Partadiyasa	387	60
Esti	527	85
Ike	415	63
Rojali	512	74

Tabel Data Bantu

4. Menghitung X^2, Y^2, XY dan total dari masing-masingnya

	X	X^2	Y	Y^2	XY
	530	280900	89	7921	47170
	300	90000	48	2304	14400
	358	128164	56	3136	20048
	510	260100	72	5184	36720
	302	91204	54	2916	16308
	300	90000	42	1764	12600
	387	149769	60	3600	23220
	527	277729	85	7225	44795
	415	172225	63	3969	26145
	512	262144	74	5476	37888
∑	4141	1802235	643	43495	279294

Mencari Koefisien “a”

5. Menghitung a dan b menggunakan rumus yang telah ditentukan

$$a = \frac{(\sum Y_i)(\sum X_i^2) - (\sum X_i)(\sum X_i Y_i)}{n \sum X_i^2 - (\sum X_i)^2}$$

$$= \frac{(643)(1802235) - (4141)(279294)}{10(1802235) - (4141)^2} = \frac{2280651}{874469} \cong 2{,}608$$

Mencari Koefisien “b”

$$b = \frac{n\ (\sum X_i\ Y_i) - (\sum X_i)(\sum Y_i)}{n \sum X_i^{\ 2} - (\sum X_i)^2}$$

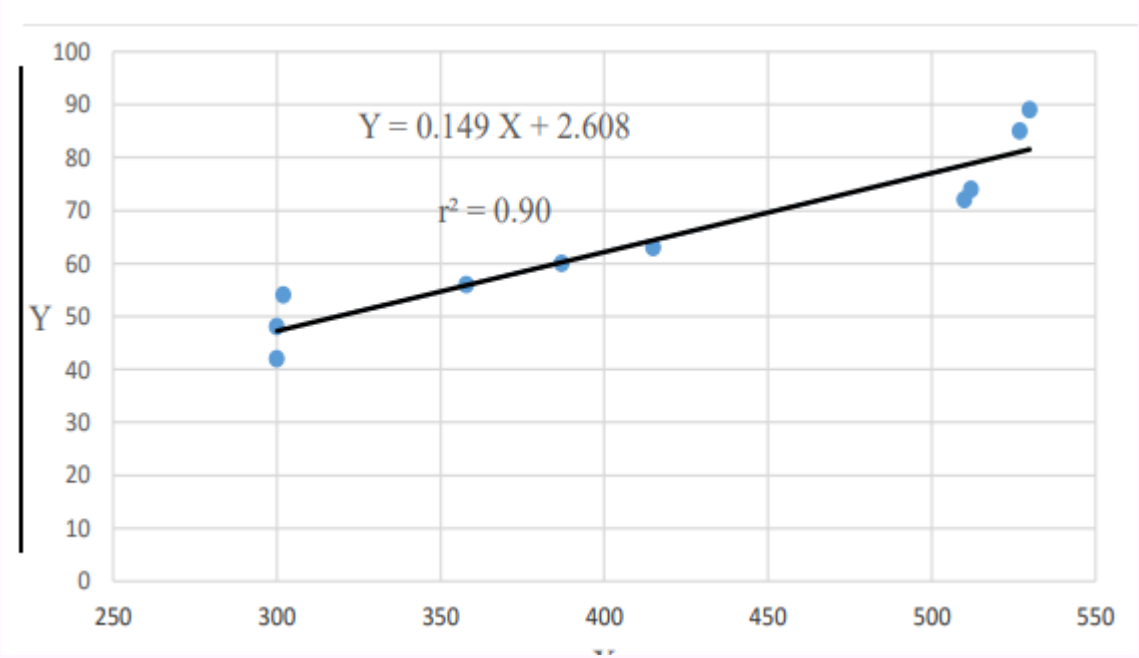
$$= \frac{10(279294) - (4141)(643)}{10(1802235) - (4141)^2} = \frac{130227}{874469} \cong 0{,}14892 \approx 0{,}149$$

Persamaan Garis Regresi

6. Membuat model Persamaan Garis Regresi

- Setelah didapat koefisien a dan b, maka persamaan garisnya adalah:

Y = 2,608 + 0,149 X



Prediksi atau Peramalan terhadap Variabel Faktor Penyebab

7. Melakukan prediksi terhadap variabel *predictor* atau *response*

- Prediksikan berat badan mahasiswa jika asupannya adalah 600 kalori/ hari:
  - Y = 2,608 + 0,149 X
  - Prediksi Y = 2,608 + (0,149 * 600) = 92 kilo gram
- Prediksikan asupan mahasiswa, jika berat badan mahasiswa adalah 40 kilo gram:
  - 40 = 2,608 + 0,149 X
  - 37,392 = 0,149X
  - X = 250.59 kalori/ hari

Uji Korelasi

- Untuk mengukur ***kekuatan hubungan antar variable predictor X dan response Y***, dilakukan analisis korelasi yang hasilnya dinyatakan oleh suatu bilangan yang dikenal dengan *koefisien korelasi*.
- Biasanya analisis regresi sering dilakukan bersama-sama dengan analisis korelasi. Persamaan koefisien korelasi (r) diekspresikan oleh:

$$r = \frac{n\sum_{i=1}^{n} Xi\,Yi - \left(\sum_{i=1}^{n} Xi\right)\left(\sum_{i=1}^{n} Yi\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} Xi\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Yi\right)^2\right]}}$$

Uji Korelasi (2)

- Berdasarkan data tabel, maka koefisien korelasinya adalah:

$$r = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}}$$

$$= \frac{10(279294) - (4141)(643)}{\sqrt{\left[10(1802235) - (4141)^2\right]\left[10(43495) - (643)^2\right]}}$$

$$= \frac{130277}{137120{,}2318} = 0{,}95$$

- Nilai ini memberi arti bahwa, hubungan variable bebas/predictor X dengan variabel terikat/ response Y adalah sangat kuat, persentasenya 95%.
- Jadi, berat badan memang sangat dipengaruhi oleh konsumsi jumlah kalori/hari

Koefisien Determinasi (r^2)

- Koefisien determinasi dapat ditentukan dengan mengkuadratkan koefisien korelasi.
- Dari contoh kasus di atas, maka koefisien determinasinya adalah $r^2 = 0{,}90$.
- Nilai ini berarti bahwa, 90% variabel bebas/ predictor X ***dapat menerangkan/ menjelaskan*** variabel tak bebas/ response Y dan 10% dijelaskan oleh variabel lainnya.

Praktik Regresi Linear (Jupyter)

In [2]:
```
# imports
import pandas as pd
import matplotlib.pyplot as plt

# this allows plots to appear directly in the notebook
%matplotlib inline
```

In [2]:
```
# read data into a DataFrame
data = pd.read_csv('http://www-bcf.usc.edu/~gareth/ISL/Advertising.csv', index_col=0)
data.head()
```

Out[2]:

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9

- Variabel bebas/predictor?
  - TV: iklan (dolar) yang dihabiskan di TV untuk satu produk di pasar tertentu (dalam ribuan dolar)
  - Radio: uang iklan yang dihabiskan untuk Radio
  - Koran: dolar iklan dihabiskan untuk Koran
- Variabel terikat/response?
  - Penjualan: penjualan satu produk di pasar tertentu (dalam ribuan gadget)

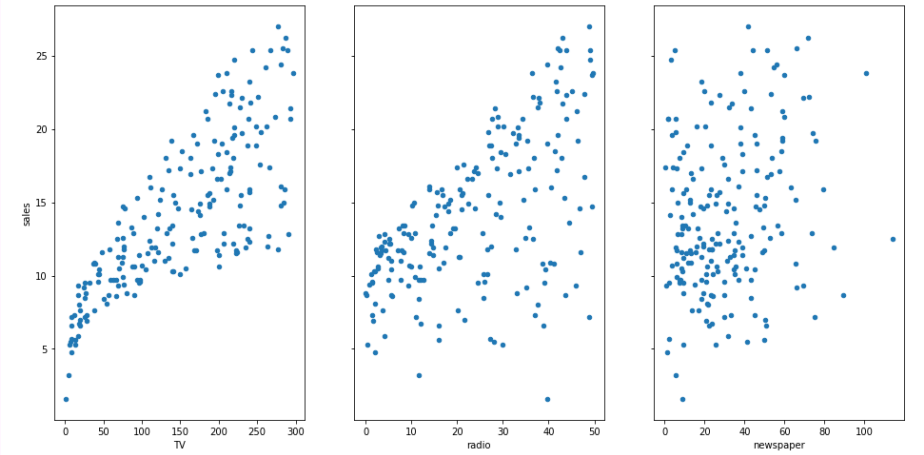
Praktik Regresi Linear (Jupyter)

```
In [3]: # print the shape of the DataFrame
        data.shape

Out[3]: (200, 4)
```

- Ada 200 pengamatan, dan dengan ada demikian 200 pasar dalam dataset.

```
In [9]: fig, axs = plt.subplots(1, 3, sharey=True)
        data.plot(kind='scatter', x='TV', y='sales', ax=axs[0], figsize=(16, 8))
        data.plot(kind='scatter', x='radio', y='sales', ax=axs[1])
        data.plot(kind='scatter', x='newspaper', y='sales', ax=axs[2])
```



Praktik Regresi Linear (Jupyter)

```
In [5]: # this is the standard import if you're using "formula notation" (similar to R)
        import statsmodels.formula.api as smf

        # create a fitted model in one line
        lm = smf.ols(formula='Sales ~ TV', data=data).fit()

        # print the coefficients
        lm.params

Out[5]: Intercept    7.032594
        TV           0.047537
        dtype: float64
```

- Bagaimana kita menginterpretasikan koefisien TV ($\beta 1$)?
  - Peningkatan "unit" dalam pengeluaran iklan di TV berkaiitan dengan kenaikan 0,047537 "unit" dalam penjualan.
  - Atau lebih jelas: Setiap penambahan $ 1.000 yang dihabiskan untuk iklan di TV berkaitan dengan peningkatan penjualan 47.537 unit TV.

Praktik Regresi Linear (Jupyter)

- Menggunakan Model untuk Prediksi:
  - Katakanlah ada pasar baru di mana pengeluaran iklan TV adalah $ 50,000. Apa yang akan diprediksi untuk Penjualan di pasar itu?

$$y = \beta_0 + \beta_1 x$$
$$y = 7.032594 + 0.047537 \times 50$$

```
In [6]: # manually calculate the prediction
        7.032594 + 0.047537*50

Out[6]: 9.409444
```

- Dengan demikian, diperkirakan Penjualan sebanyak 9409 unit di pasar itu.

Praktik Regresi Linear (Jupyter)

- Prediksi menggunakan statsmodel:

```
In [7]: # you have to create a DataFrame since the Statsmodels formula interface expects it
        X_new = pd.DataFrame({'TV': [50]})
        X_new.head()
```

Out[7]:

	TV
0	50

```
In [8]: # use the model to make predictions on a new value
        lm.predict(X_new)
```

```
Out[8]: array([ 9.40942557])
```

- Menghitung r squared:

```
In [14]: # print the R-squared value for the model
         lm.rsquared
```

```
Out[14]: 0.61187505085007099
```

Praktik Regresi Linear (Jupyter)

- Plotting the Least Squares Line

In [9]:
```
# create a DataFrame with the minimum and maximum values of TV
X_new = pd.DataFrame({'TV': [data.TV.min(), data.TV.max()]})
X_new.head()
```

Out[9]:

	TV
0	0.7
1	296.4

In [10]:
```
# make predictions for those x values and store them
preds = lm.predict(X_new)
preds
```

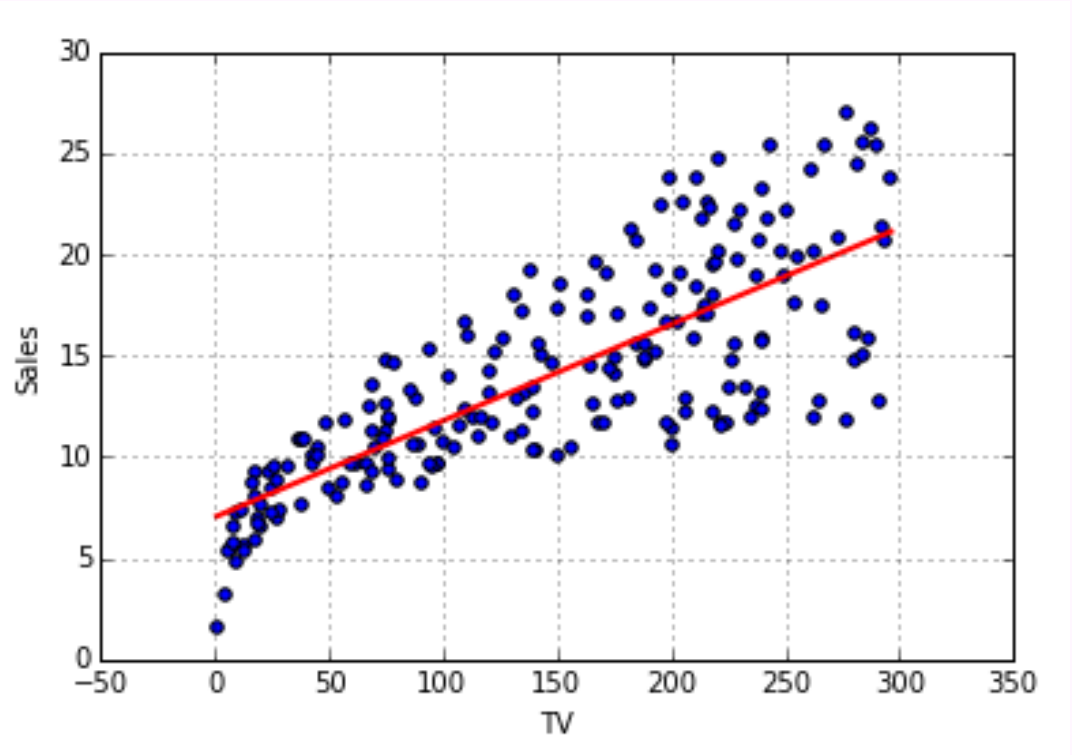
Out[10]:
```
array([  7.0658692 ,  21.12245377])
```

In [11]:
```
# first, plot the observed data
data.plot(kind='scatter', x='TV', y='Sales')

# then, plot the least squares line
plt.plot(X_new, preds, c='red', linewidth=2)
```

Praktik Regresi Linear (Jupyter)

- Plotting the Least Squares Line



- Referensi:
- http://www-bcf.usc.edu/~gareth/ISL/
- http://people.duke.edu/~rnau/regintro.htm#history
- https://towardsdatascience.com/introduction-to-linear-regression-in-python-c12a072bedf0?gi=9e4b95a7604f
- https://raw.githubusercontent.com/justmarkham/DAT4/master/notebooks/08_linear_regression.ipynb