

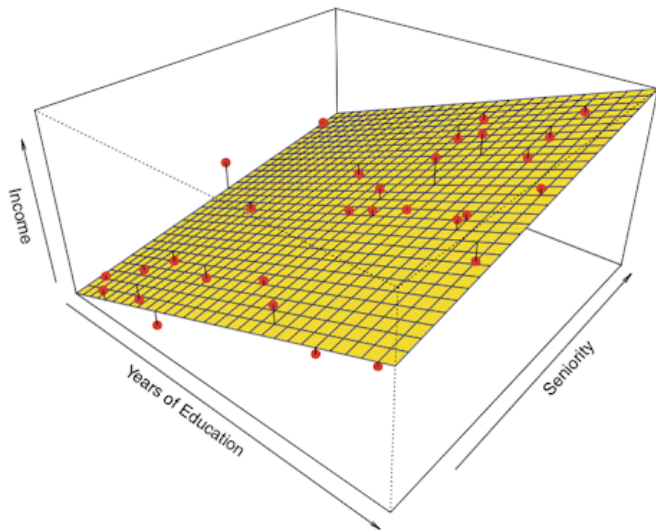
Introduction to data science & artificial intelligence (IF7100)

Arthur Charpentier

#322 Multiple Regression

été 2020

Linear Regression



From 1 to k features

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \varepsilon_i \text{ or } y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i,$$

From a mathematical perspective, use matrix notations

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_{\mathbf{y}, n \times 1} = \underbrace{\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{pmatrix}}_{\mathbf{X}, n \times (k+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}, (k+1) \times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}, n \times 1}.$$

i.e. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Assume that $\mathbb{E}(\varepsilon_i) = 0$ and $\text{Var}[\varepsilon_i] = \sigma^2$.

The OLS estimator is $\hat{\boldsymbol{\beta}} \in \text{argmin} \sum_{i=1}^n (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2$, i.e.

$$\hat{\boldsymbol{\beta}} = \text{argmin} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Linear Regression

Then $\mathbb{E}(\hat{\beta}) = \beta$ and $\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$.

Assuming either that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ or n large,

$$\hat{\beta} \approx \mathcal{N}(\beta, \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1})$$

As previously (#321) we can test for significance $H_0 : \beta_j = 0$ and we can derive some confidence intervals, for β_j (for any j)

For prediction, $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_k x_k$.

Ceteris paribus can be translated into "all other things being equal" or "holding other factors constant"

Mutatis mutandis approximately translates as "allowing other things to change accordingly" or "the necessary changes having been made"

Regression

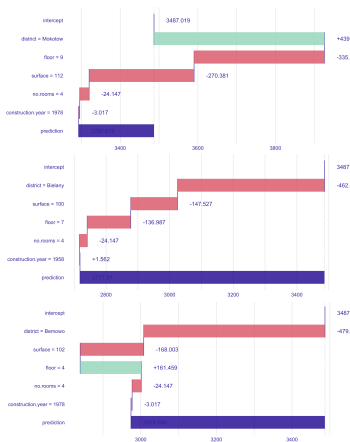
```
1 > import numpy as np
2 > import statsmodels.api as sm
3 > x = np.array([[5,1], [15,4], [25,-5], [35,4],
4               [45,-2], [55,2]])
5 > x = sm.add_constant(x)
6 > y = np.array([5, 20, 14, 32, 22, 38])
7 > model = sm.OLS(y, x)
8 > results = model.fit()
9 > print(results.summary())
```

```
=====
              coef.  std err          t      P>|t|      [0.025      0.975]
-----
12 const          4.0581    3.370     1.204    0.315    -6.668    14.785
13 x1              0.5578    0.097     5.770    0.010     0.250     0.865
14 x2              1.5604    0.508     3.071    0.055    -0.057     3.178
15 -----
16 Dep. Variable:    y              R-squared:            0.931
17 Model:              OLS          Adj. R-squared:       0.886
18                               F-statistic:           20.37
```

Regression

```
1 > df = data.frame(x1 = c(5, 15, 25, 35, 45, 55),
2                   x2 = c(1, 4, -5, 4, -2, 2),
3                   y = c(5, 20, 14, 32, 22, 38))
4 > model = lm(y~x1+x2, data=df)
5 > summary(model)
6
7 Coefficients:
8             Estimate Std. Error t value Pr(>|t|)
9 (Intercept)  4.05810    3.37049   1.204   0.3149
10 x1           0.55783    0.09667   5.770   0.0103 *
11 x2           1.56037    0.50818   3.071   0.0545 .
12 ---
13 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
14
15 Residual standard error: 4.037 on 3 degrees of freedom
16 Multiple R-squared:  0.9314, Adjusted R-squared:  0.8857
17 F-statistic: 20.37 on 2 and 3 DF,  p-value: 0.01796
```

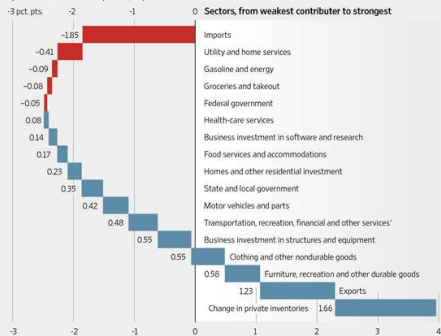
Break-down for Additive Models



The Path to Growth | How GDP came to grow at a 4.0% pace

While an increase in imports subtracted from growth in gross domestic product, a buildup of business inventories and an increase in consumer spending more than offset the headwinds.

Percentage-point contribution to annualized U.S. GDP growth in the second quarter, by select sector and component, adjusted for inflation and seasonal variations



*Includes insurance services and final consumption expenditures of nonprofit institutions serving households
 Note: Percentages don't total to complete GDP movement due to minor differences in rounding and measurement
 Source: Commerce Department

Andrew Van Dam/The Wall Street Journal

See [The Wall Street Journal](#), on accounting equations

Partial Dependence Plot

Introduced in **Greedy function approximation: A gradient boosting machine**, Friedman (2001)

Let \mathbf{x} be split in two parts : \mathbf{x}_s (variable(s) of interest) and \mathbf{x}_c the complementary, $\mathbf{x} = (\mathbf{x}_s, \mathbf{x}_c)$. Partial dependence of \mathbf{x}_s is

$$p(\mathbf{x}_s) = \mathbb{E}[m(\mathbf{x}_s, \mathbf{S}_c)] \text{ and } \hat{p}(\mathbf{x}_s) = \frac{1}{n} \sum_{i=1}^n \hat{m}(\mathbf{x}_s, \mathbf{x}_{i,c})$$

