Solving the monster of collinearity

By Alejandro Mahecha
June 25 2020

1. Definition and collinearity examples.
2. Synthetic data simulation OLS
3. Simulation Ridge, Lasso and elastic-net.
4. Bayesian regression
5. Conclusions and next steps

Multicollinearity

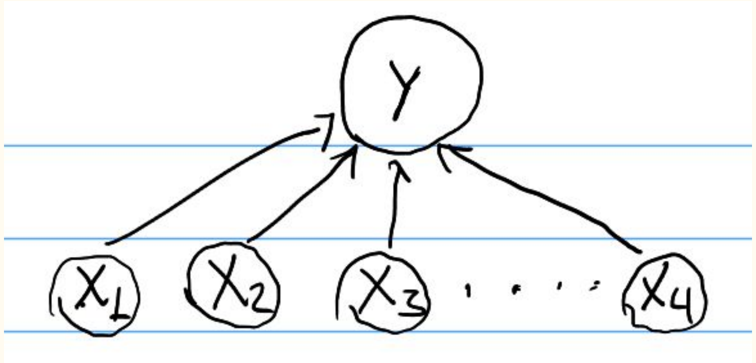
volume_sales(price, %_diff_price, ...)

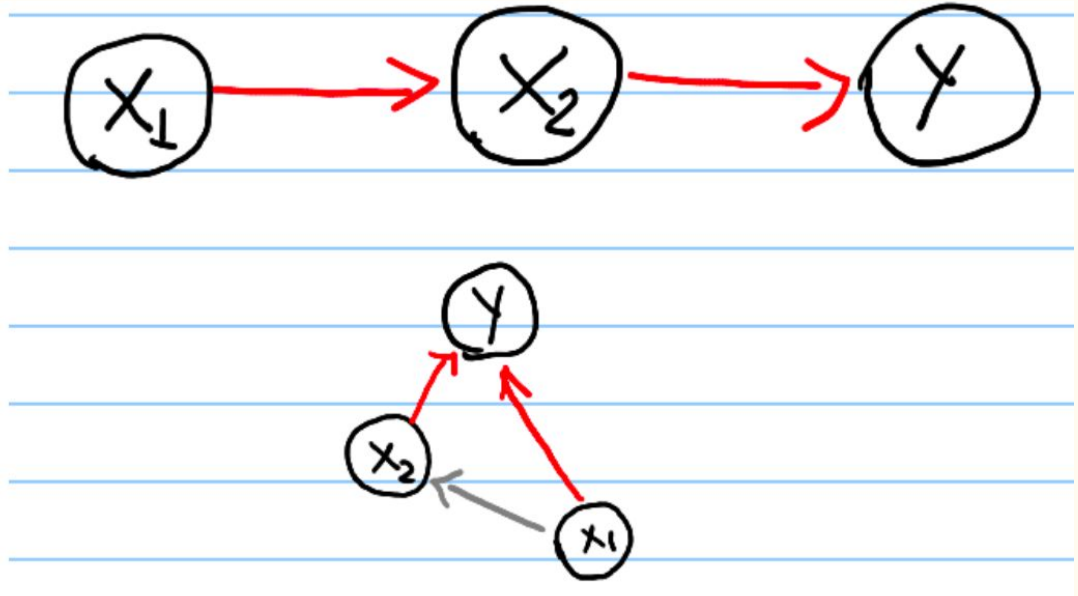
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

$$\downarrow \beta_1 , \beta_2 \uparrow$$

salary(age, years_experience, ...)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$


Y
X1
X2
X3
X4


X1
X2
Y
Y
x2
x1

$$x_2 = x_1 + \epsilon$$

$$x_1, \epsilon \sim \mathcal{N}(\mu, \sigma^2).$$

$$y = x_1 + x_2$$

```
mu, sigma = 2, 0.9
x1 = np.random.normal(mu, sigma, size=100)
x2 = np.random.normal(loc=x1, scale=0.05)
y = np.random.normal(loc=x1+x2, scale=1)
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.816
Model:	OLS	Adj. R-squared:	0.812
Method:	Least Squares	F-statistic:	214.6
Date:	Fri, 29 May 2020	Prob (F-statistic):	2.41e-36
Time:	16:13:19	Log-Likelihood:	-128.01
No. Observations:	100	AIC:	262.0
Df Residuals:	97	BIC:	269.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.2954	0.216	-1.365	0.176	-0.725	0.134
x1	3.6637	1.868	1.962	0.053	-0.043	7.371
x2	-1.5003	1.878	-0.799	0.426	-5.228	2.227

Assumptions

- Predictors have to be independent(pearson corr).
- Probability distributions of predictors has influence on coefficients precision.

$$(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \dfrac{1}{1-r_{12}^2} & \dfrac{-r_{12}}{1-r_{12}^2} \\ \dfrac{-r_{12}}{1-r_{12}^2} & \dfrac{1}{1-r_{12}^2} \end{bmatrix}$$

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1-r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1-r_{12}^2}$$

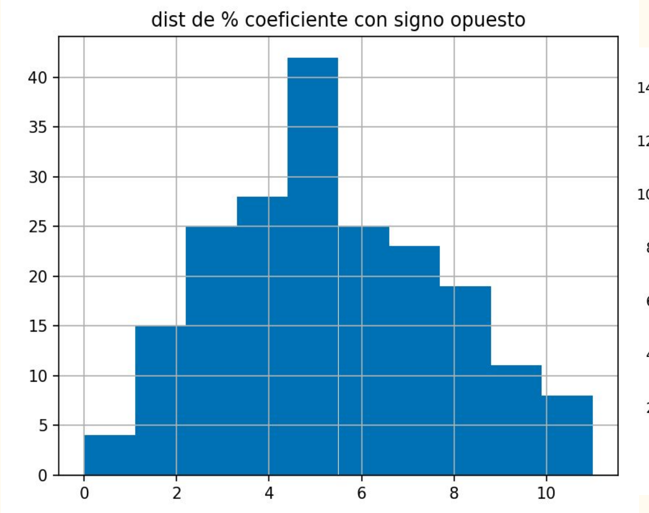
$$|r_{12}| \to 1, \ \mathrm{Var}\left(\hat{\beta}_j\right) = C_{jj}\sigma^2 \to \infty$$

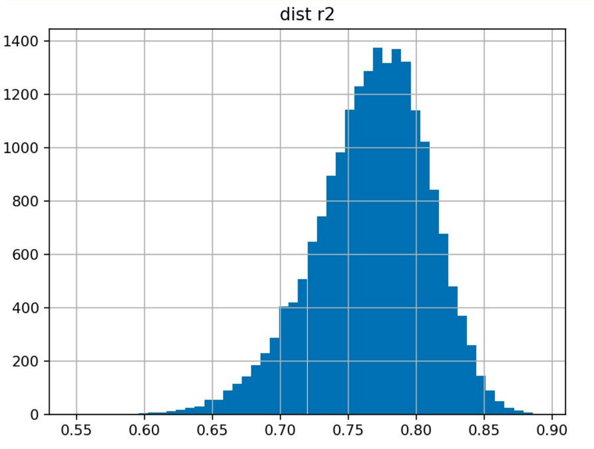
Simulations

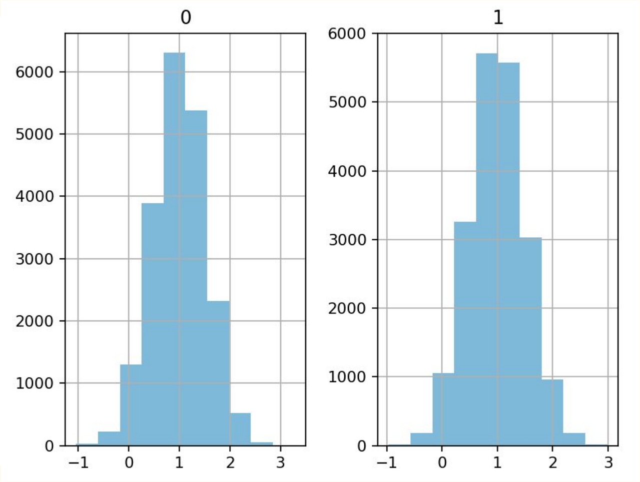
Questions

- % opposite sign coeff dist?
- R_2 dist?
- Coeffs dists?

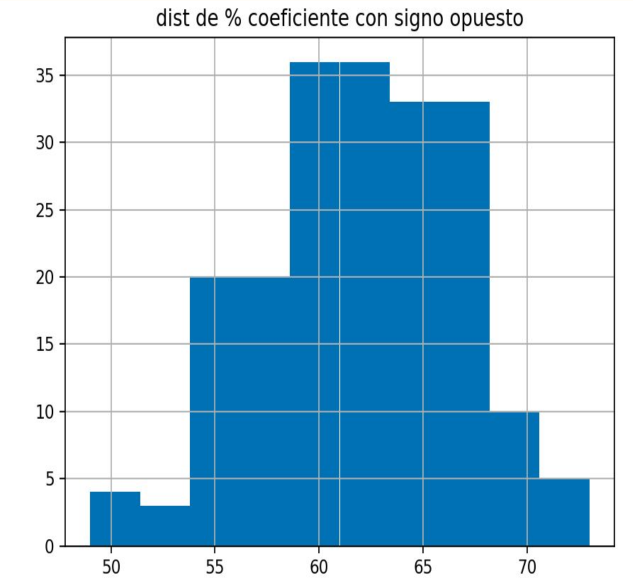
Low collinearity $\sigma = 0.2$

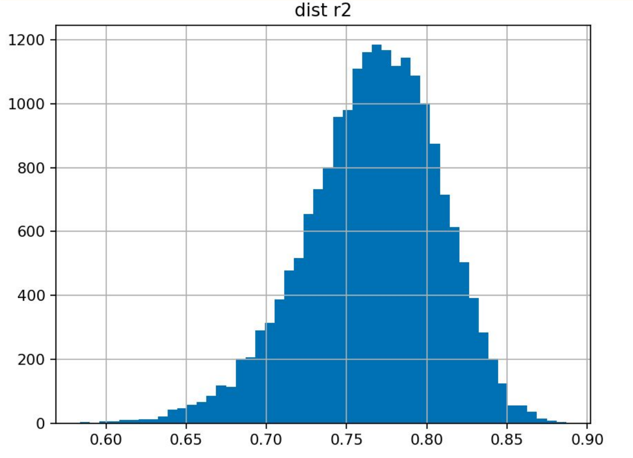


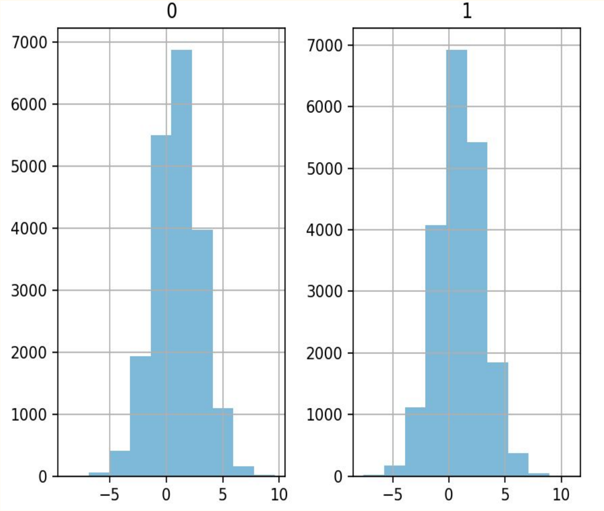




High collinearity $\sigma = 0.05$





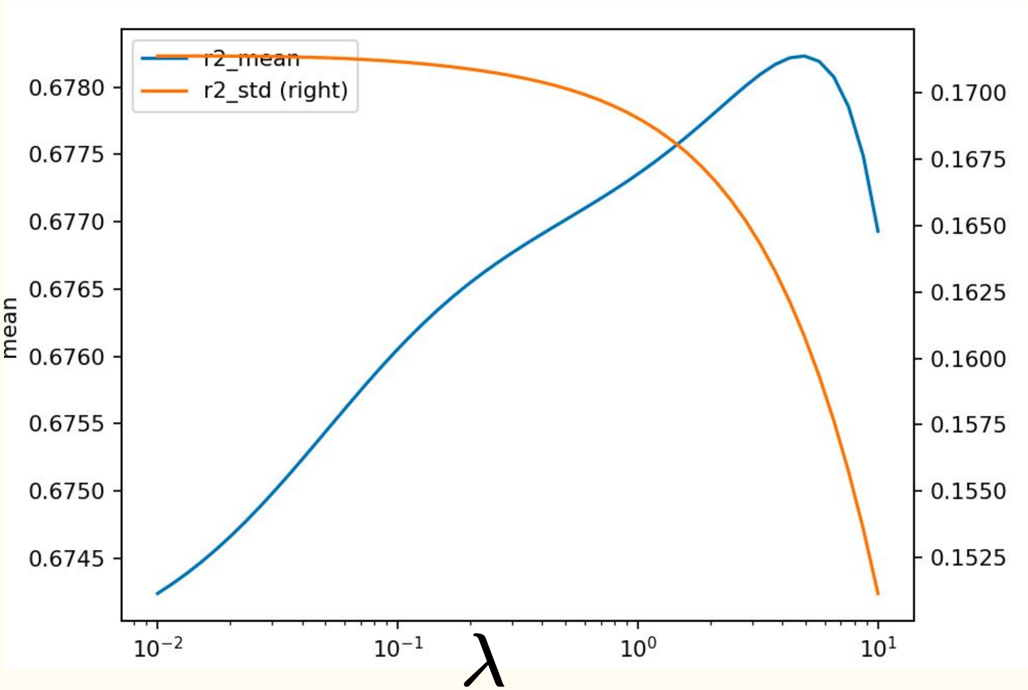


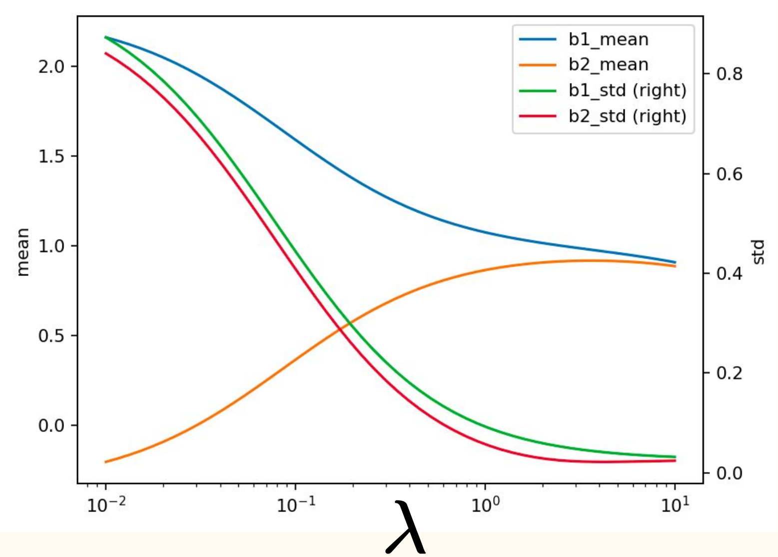
Regularization

- Ridge
- Lasso
- ElasticNet

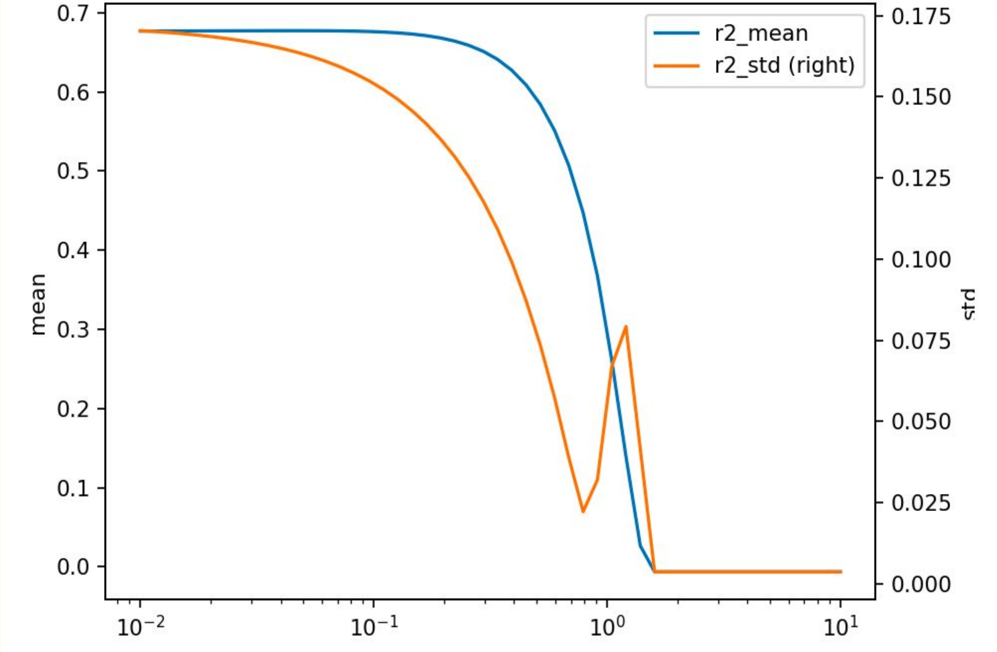
$$\text{Ridge:} \quad (\boldsymbol{Y} - \boldsymbol{X}\beta)^2 + \lambda\beta^2$$

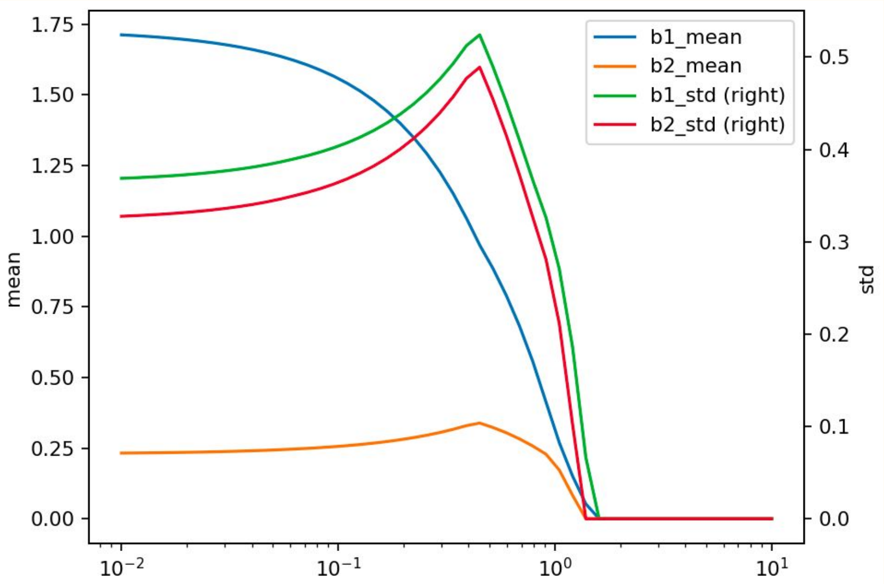
Test				
	Test			
		Test		
			Test	
				Test



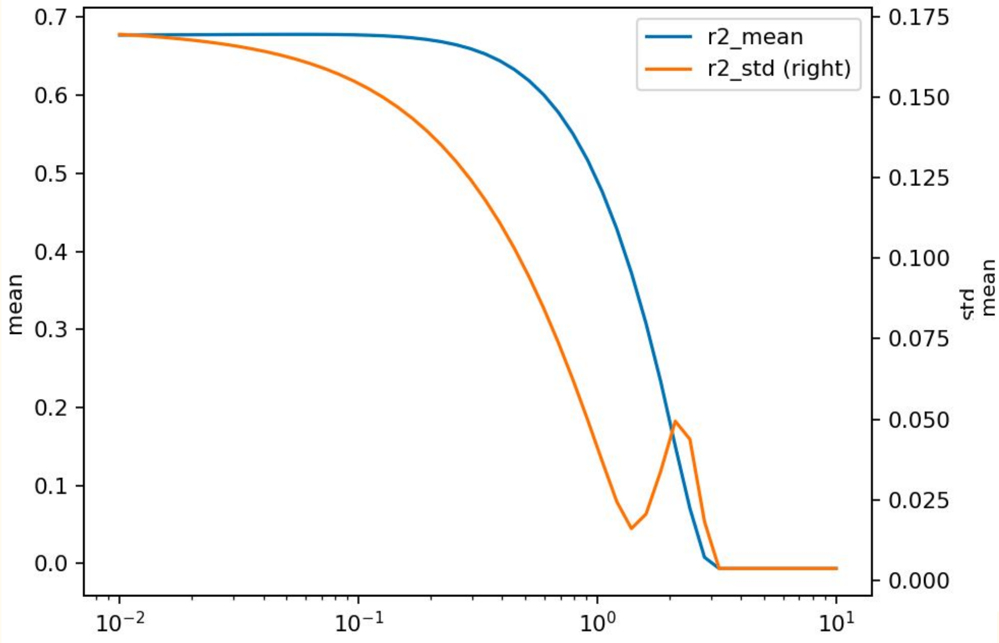


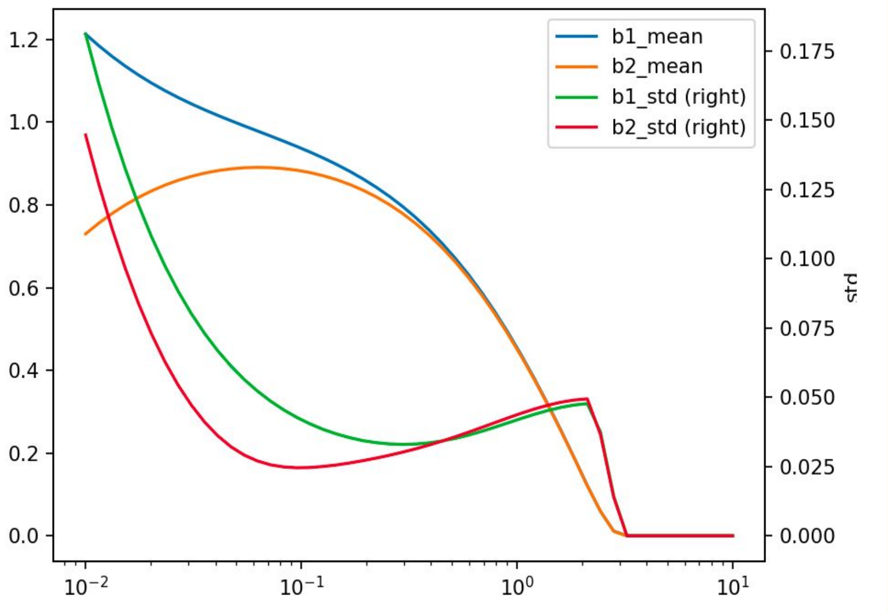
Lasso: $(\boldsymbol{Y} - \boldsymbol{X}\beta)^2 + \boldsymbol{\lambda}\,|\beta|$





ElasticNet: $(Y - X\beta)^2 + \lambda(\epsilon\,|\beta| + (1-\epsilon)\beta^2)$





Bayes Regression model

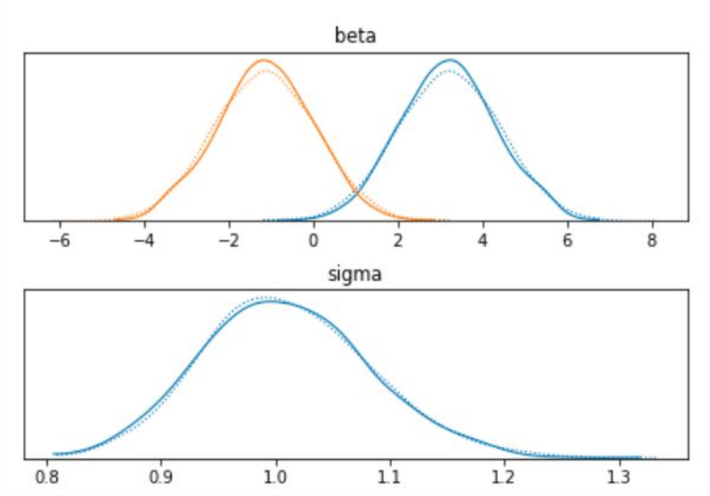
$$Y \sim \mathcal{N}(\mu, \sigma^2)$$
$$\mu = \alpha + \beta_1 X_1 + \beta_2 X_2$$

$$\pi(\theta \,|\, \mathbf{Y} = \mathbf{y}) \;=\; \frac{P(\mathbf{Y} = \mathbf{y} \,|\, \theta)\, \pi(\theta)}{P(\mathbf{Y} = \mathbf{y})} \;=\; \frac{P(\mathbf{Y} = \mathbf{y} \,|\, \theta)\, \pi(\theta)}{\int_{0,1} P(\mathbf{Y} = \mathbf{y} \,|\, \theta)\, \pi(\theta)\, d\theta}.$$

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{Y}, \mathbf{X}) \;\propto\; f_Y(\mathbf{Y} \,|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2)\, f_\beta(\boldsymbol{\beta}|\sigma^2)\, f_\sigma(\sigma^2)$$

Bayes Linear regression

Pymc3



scikit-learn

```
In [86]: reg.coef_
Out[86]: array([0.80923041, 1.25820039])

In [87]: lireg.coef_
Out[87]: array([-0.43266424,  2.50474319])
```

References

1) https://iopscience.iop.org/article/10.1088/1742-6596/1265/1/012021/pdf
2) https://arxiv.org/pdf/1509.09169.pdf