Linear regression: Regularization and Bayesian way

Agoston Torok

Multivariate statistics

ELTE

Assumptions of linear regression I.

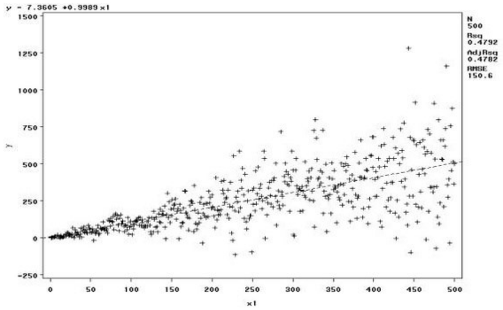
- Weak exogeneity: means that in the model we specify only ε as a random variable, *x* is error-free fixed values → Was there no ε we were able to get SS = 0

Typically too idealistic

- Linearity: all the β terms are simple summed (note you can transform or combine predictors to include more complex effects, but the model will still be linear in form)

We often DO transform predictors

- Constant variance (homoscedasticity): the error is the same for every x



Assumptions of linear regression II.

- Independence of errors of predictors: The emphasis on the errors, that is the predictors can be correlated, but their errors (which would easily violate homoscedasticity) should not be correlated

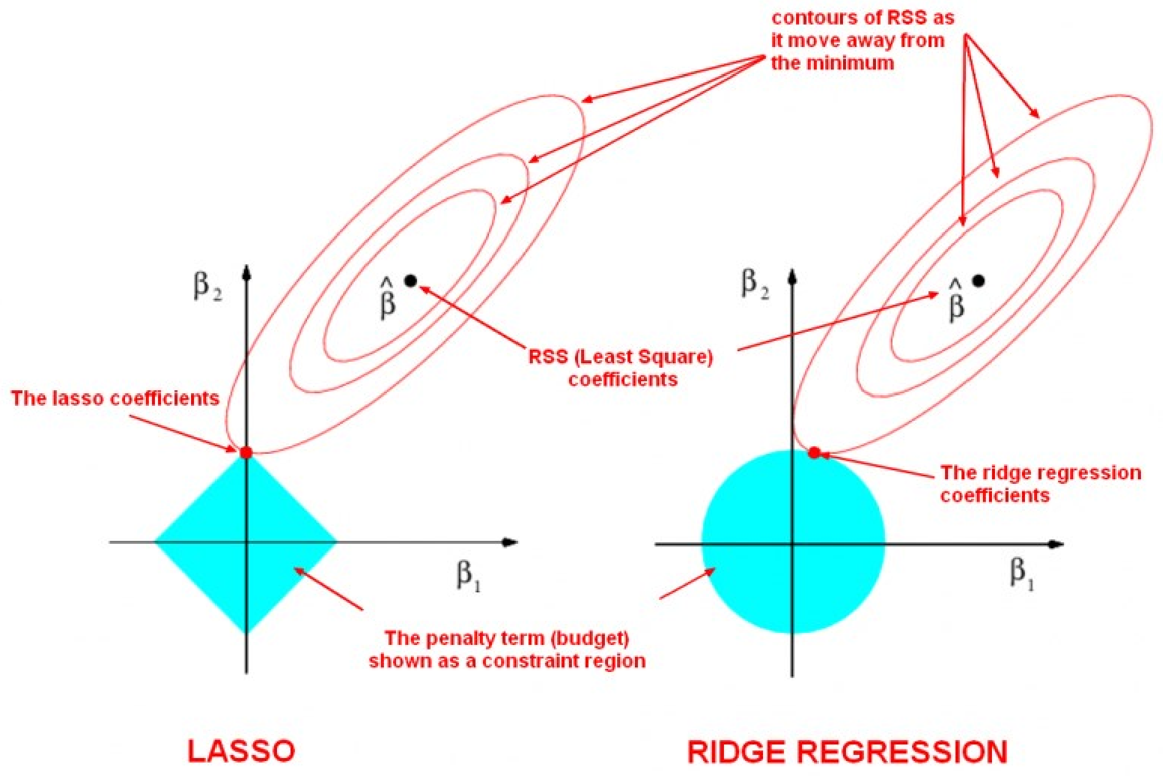
- No multicollinearity: The predictors should not be (almost) perfectly correlated. This is not necessarily bad for the model, but definetly bad for the parameter estimation

Regularization

- Assume: large number of predictors, collinearity, looking for feature selection

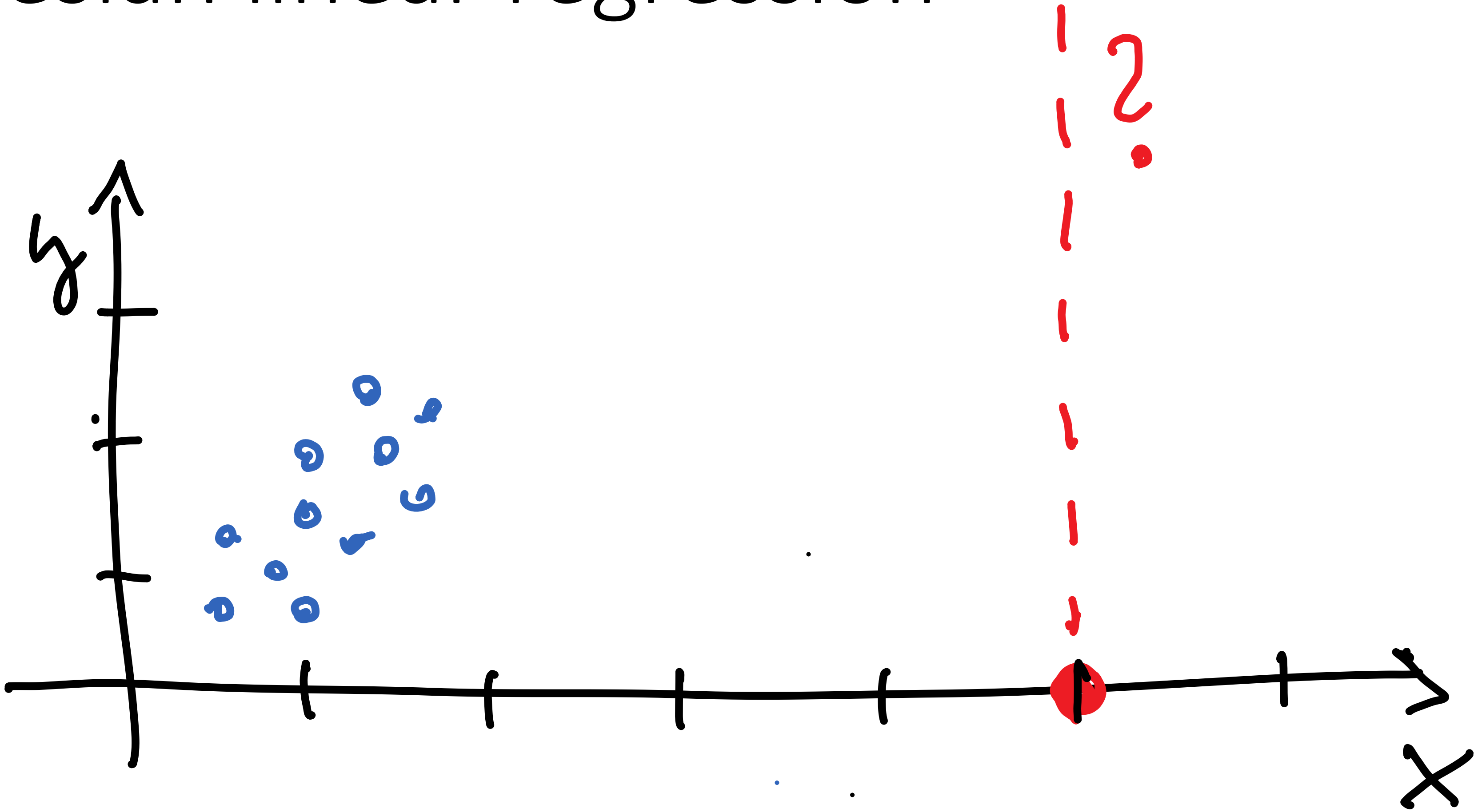
- The problem with OLS is that it tries to maximize the model fit to the data → overfitting

- Regularization prevents this by adding constraints on the model

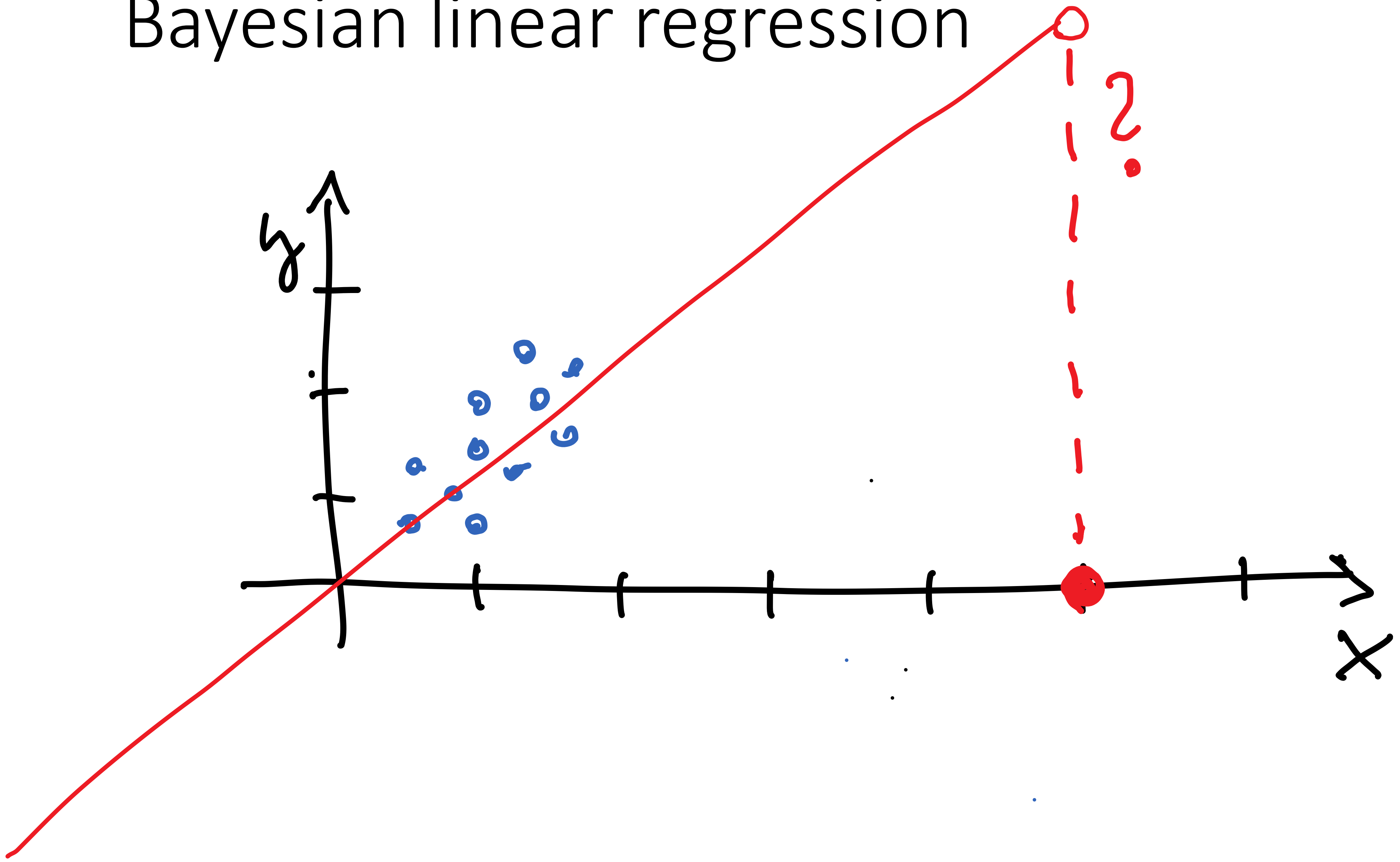


Constraints on the model …
so you actually want to include some model?

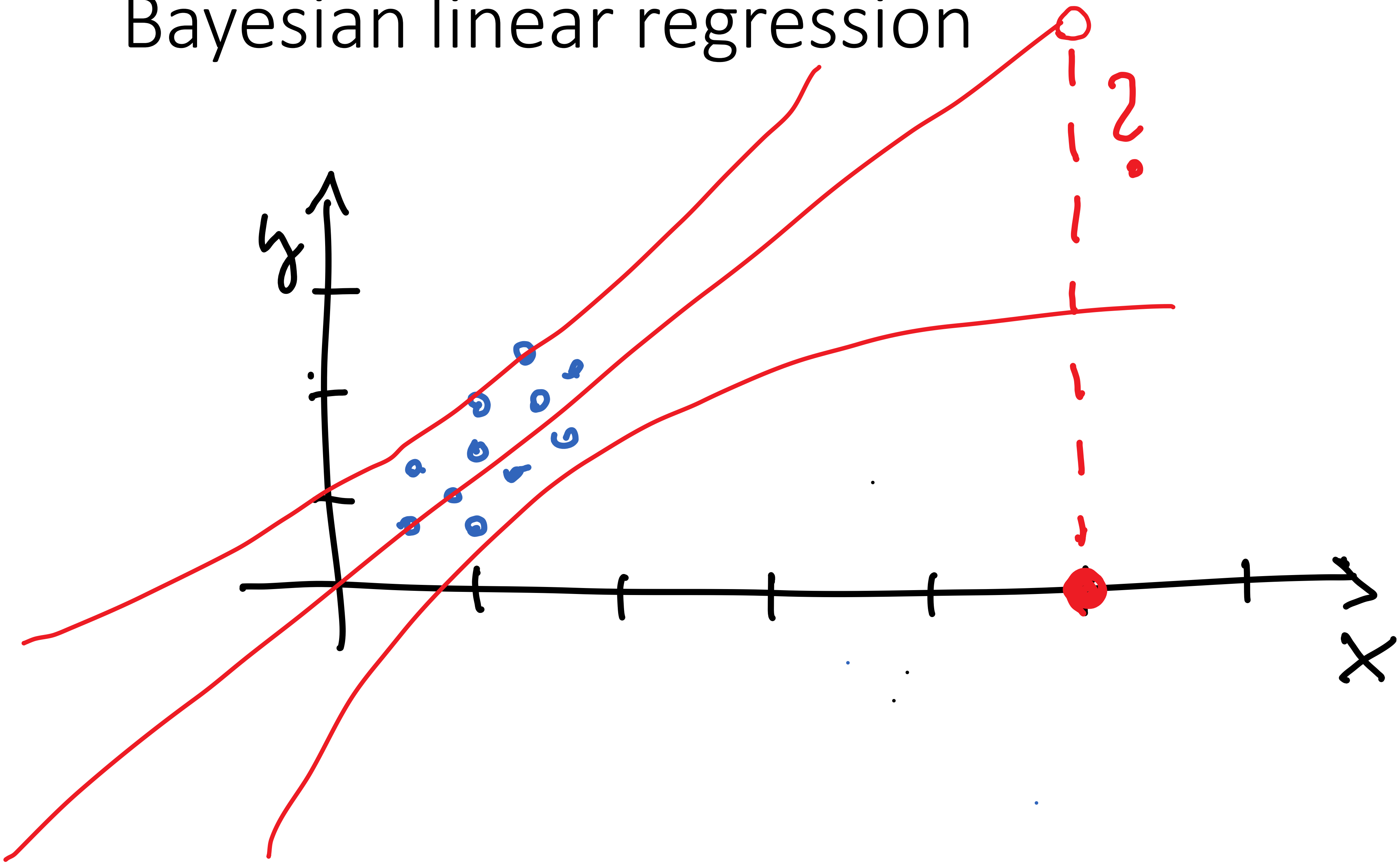
Bayesian linear regression



Bayesian linear regression

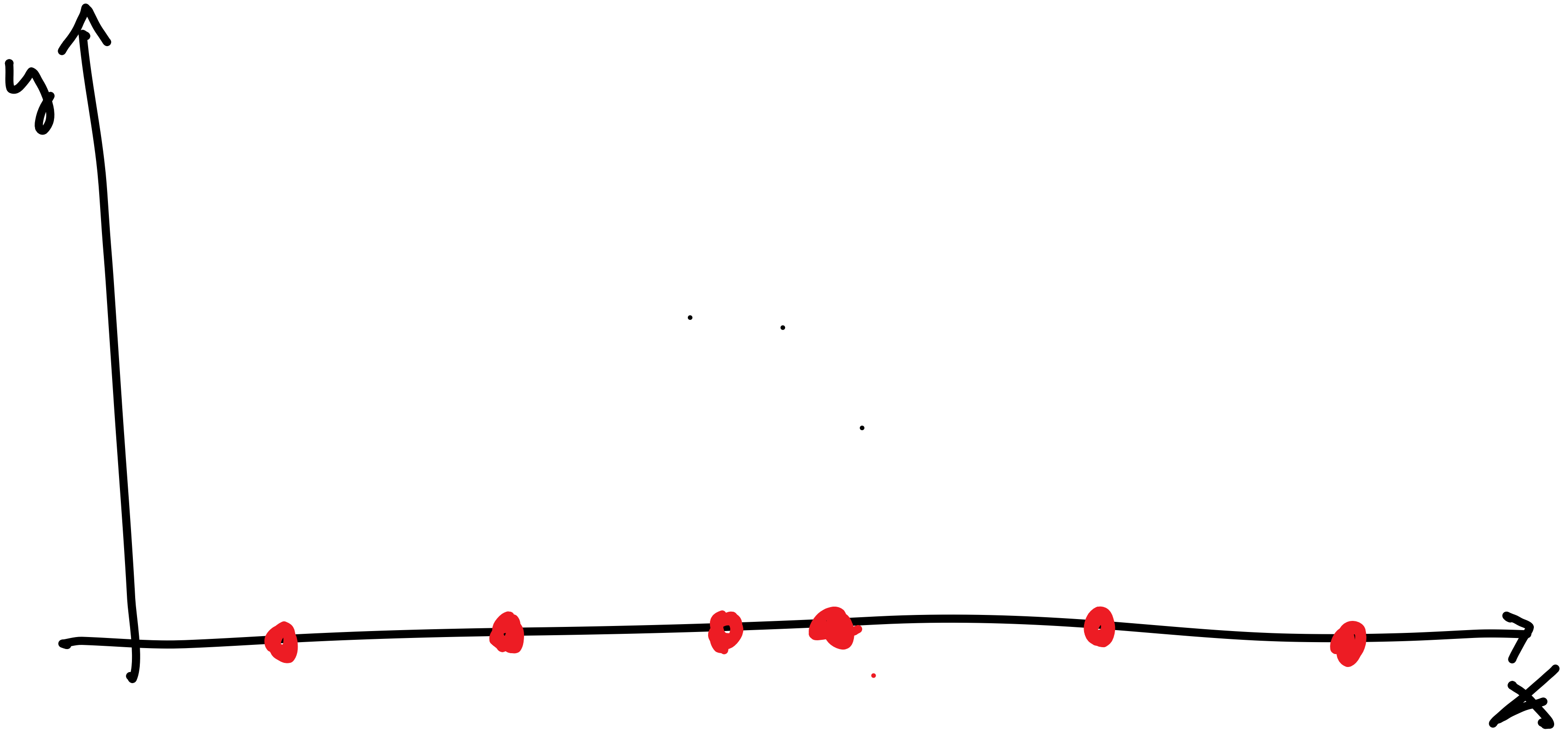


Bayesian linear regression



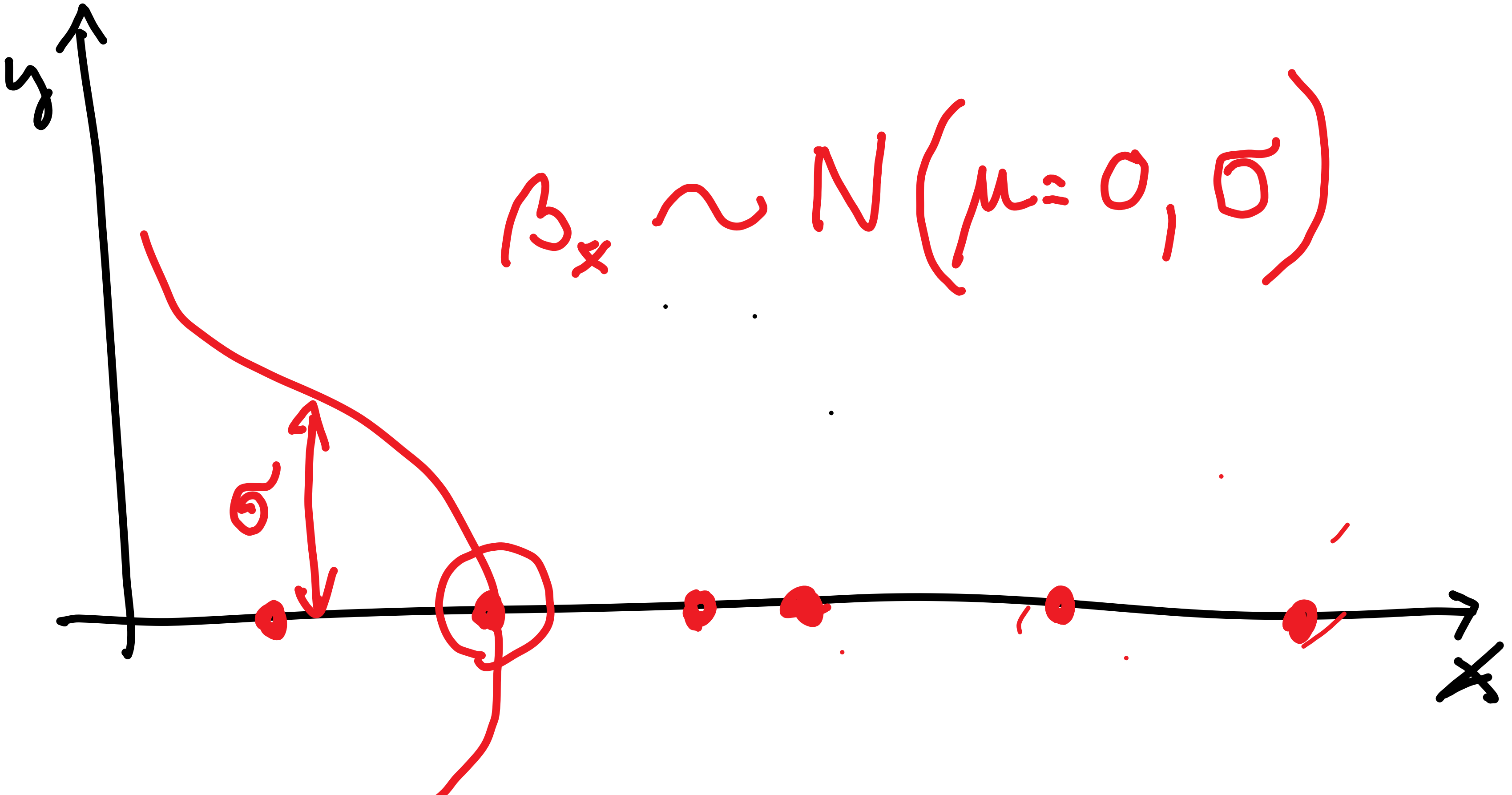
How to choose the prior?

- Objective (un-informed) vs. Subjective (informed) prior



How to choose the prior?

- Objective (un-informed) vs. Subjective (informed) prior



Important remarks

- Choosing the uninformative prior is similar to Ridge regression
- OLS – if all the assumptions are fulfilled then the estimate of parameters of the mean is good
- Bayesian LR – Conditional probabilities – P(Y|X) – makes it able to include variable uncertainty (SD of the prediction) for X values