LINEAR REGRESSION II

11.16.2018

HOMEWORK 4

* is due monday after next!

RECAP

* how do we solve linear regression problems?

  * find weights that minimize the sum of squared errors!

  * (squared error function is a PARABOLA)

  * this requires *simultaneously* estimating all the weights

RECAP

* gradient descent!

  * given the current settings of the weights, what small change would decrease the error the most?

  * take many tiny steps, this will eventually lead to the right answer

* or.. an analytic solution!

ANALYTIC REGRESSION

* there is an equation that can exactly solve the least-squared-error problem

  * (if this is what you want to do! sometimes it is, sometimes it isn’t)

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y$$

ANALYTIC REGRESSION

* **np.linalg.lstsq** solves least squares regression (example)
* it returns 4 things:
  * the regression weights (beta)
  * the residuals (final squared error)
  * the rank (we'll talk about this later)
  * the singular values (ditto)

EVALUATING REGRESSION MODELS

* how do you know if a regression model is *good*?

* one common metric is R^2, also called the **coefficient of determination** or **variance explained**

EVALUATING REGRESSION MODELS

* R^2 = 1 - (RSS / TSS)

* where RSS is the “residual sum of squares” (this is squared error, which we’ve seen)

* and TSS is the “total sum of squares” (like squared error if your model always predicted zero)

EVALUATING REGRESSION MODELS

* we can also define it in terms of variance

* ***R^2*** = 1 - (var(y-y_hat) / var(y))

* what’s the difference between squared error and variance?

EVALUATING REGRESSION MODELS

* suppose that we are given a matrix of variables (aka regressors) **X**, and a vector of outputs **Y**

* we fit a linear model **Y_hat = X . beta**

* then we evaluate it by computing $\mathbf{R^2}$ using **X** and **Y**

* what are the possible values of $\mathbf{R^2}$?

IN-SET VS. OUT-OF-SET EVALUATION

* evaluating a regression model using the same data that we used to train/estimate/fit it is called *in-set evaluation*

* in-set evaluation is biased *upward*, and the amount of bias depends on the number of regressors in the model

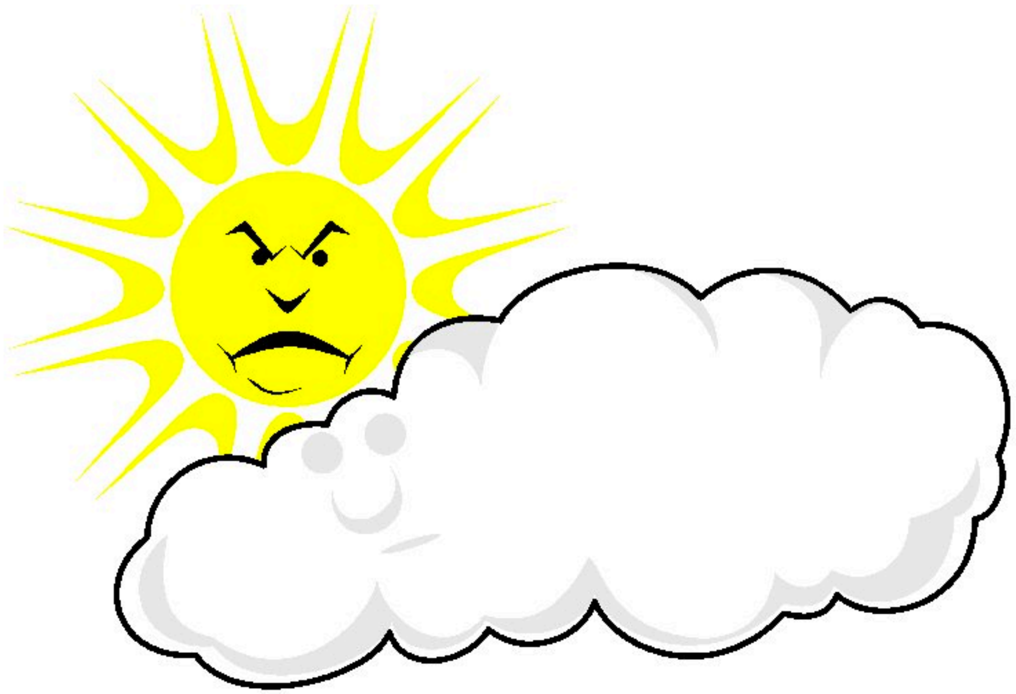
IN-SET VS. OUT-OF-SET EVALUATION

* for example: suppose we have *N* data points and *N* regressors that are pure noise—they have no relationship to the output whatsoever

* in-set variance explained is EXACTLY 1.0

* *THE MODEL IS PERFECT*

* *****THIS IS BOGUS*****



IN-SET VS. OUT-OF-SET EVALUATION

* instead, what if you split up your X and Y into “training” and “test” sets?

* you could fit your regression model using (X_trn, Y_trn), and then test how well it works on (X_test, Y_test)!

* is $\mathbf{R}^2$ biased in this case? What possible values can it take?

REGRESSION STABILITY

* as hinted at on wednesday, ordinary least squares regression has a problem:

* if two regressors are similar (i.e. correlated), then there are many possible weight combinations that would give ~the same answer!

* which set of weights is “best” ends up being totally determined by *noise* (example)

REGRESSION STABILITY

* this is bad: if your weights are essentially random, they are ~impossible to interpret, and model performance can suffer, so:

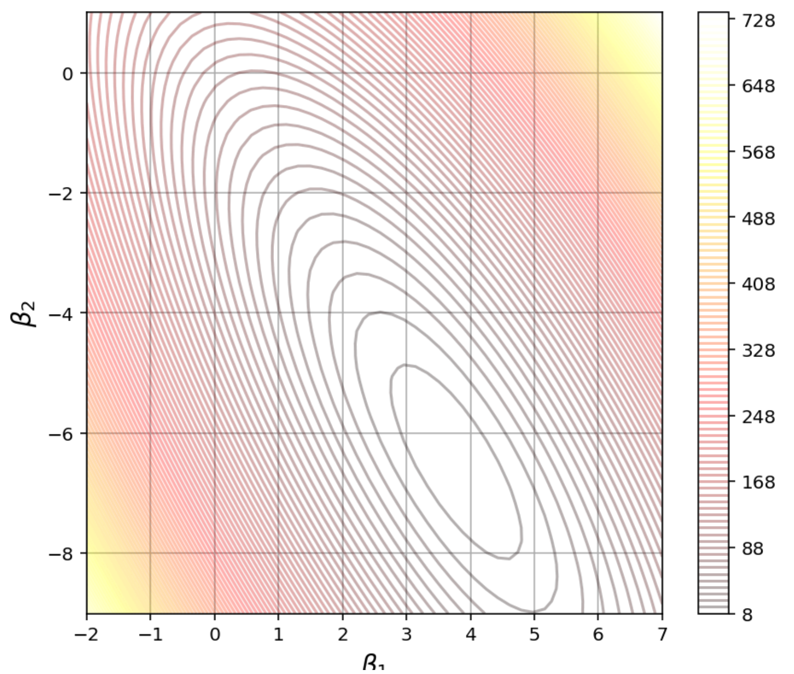
* (1) let’s figure out when this is happening, and

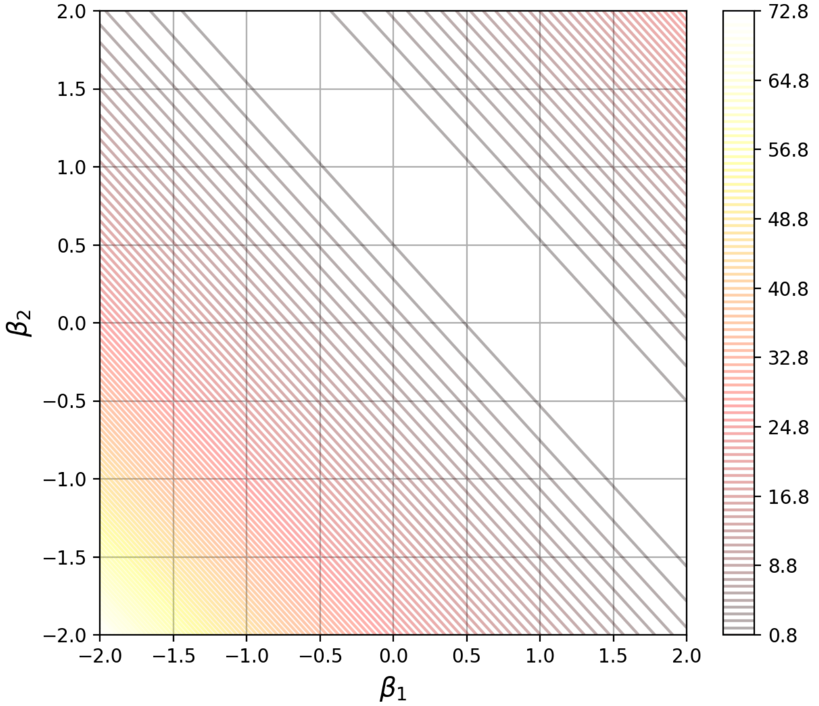
* (2) let’s stop it from happening

REGRESSION STABILITY

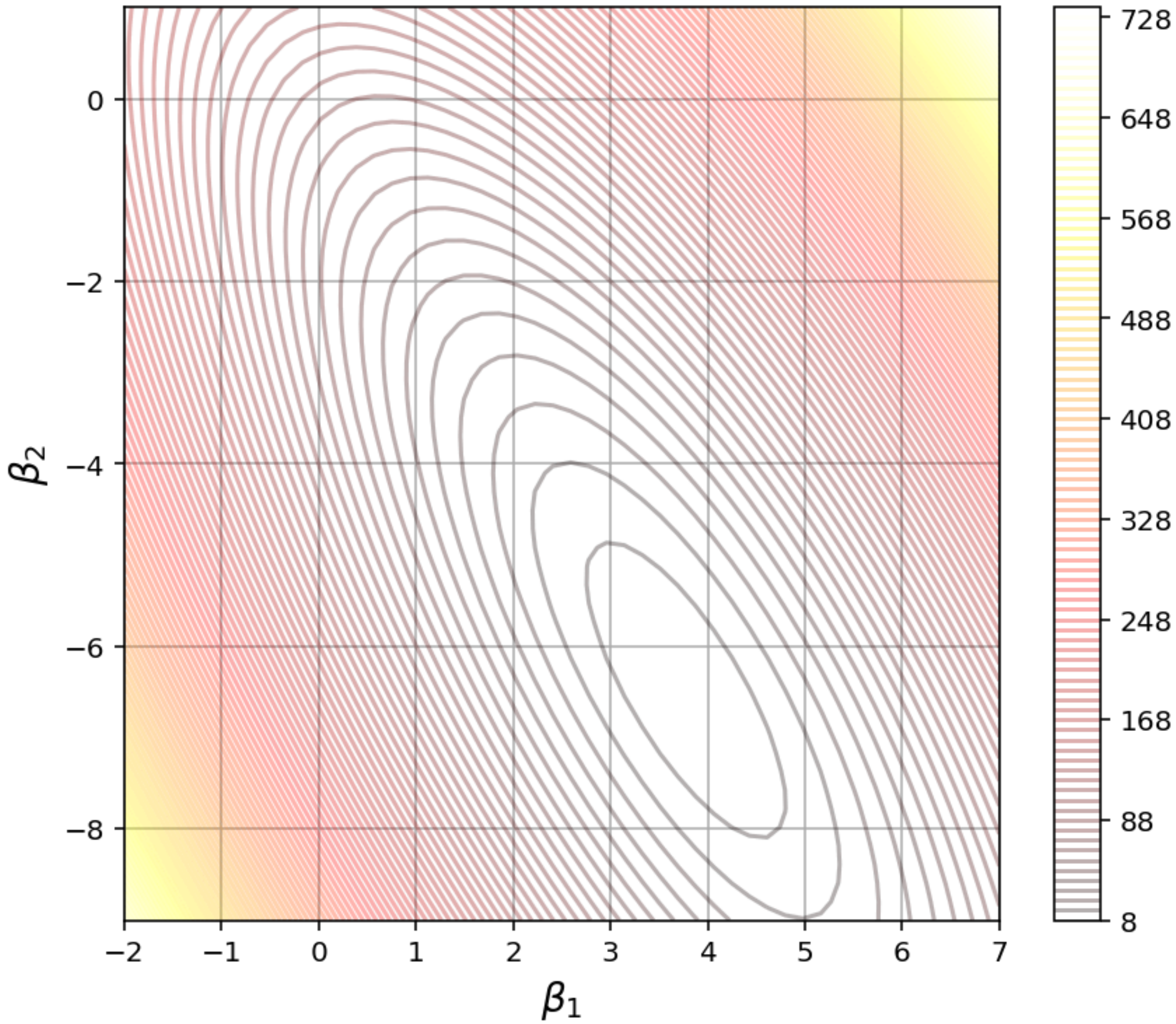
* how do we know when regression is unstable?

* it’s related to the shape of the error function!

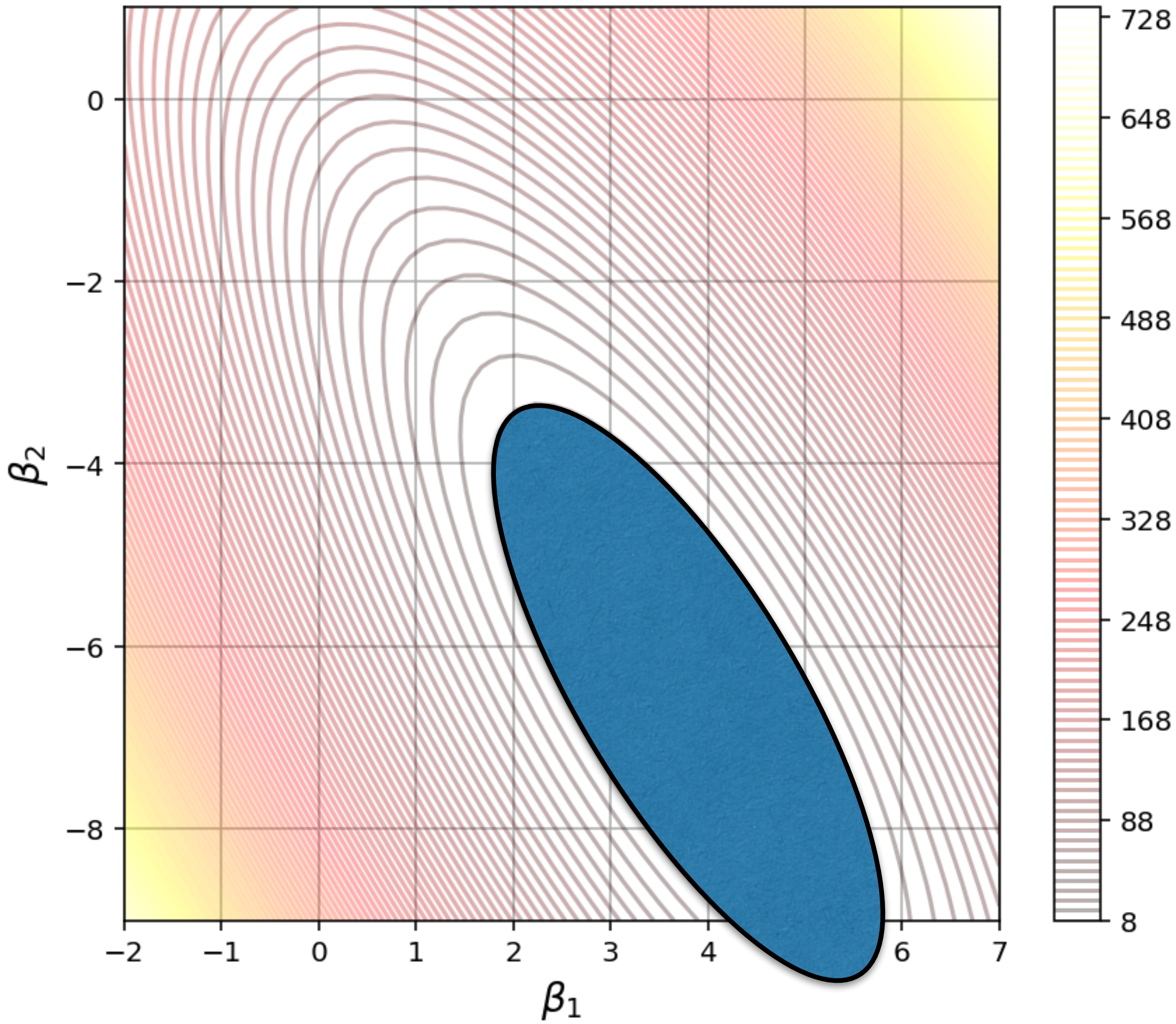




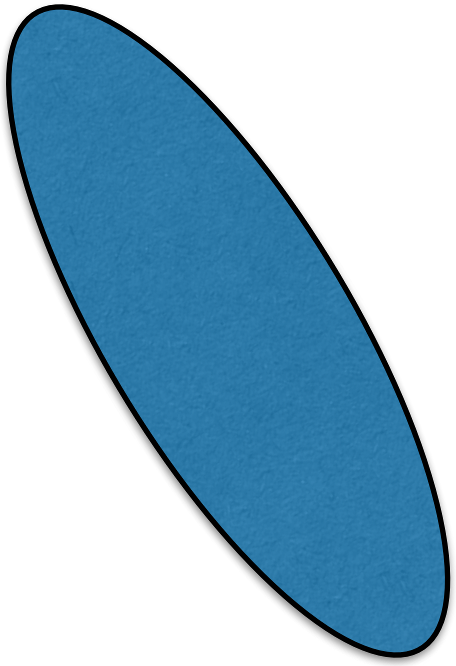
REGRESSION STABILITY



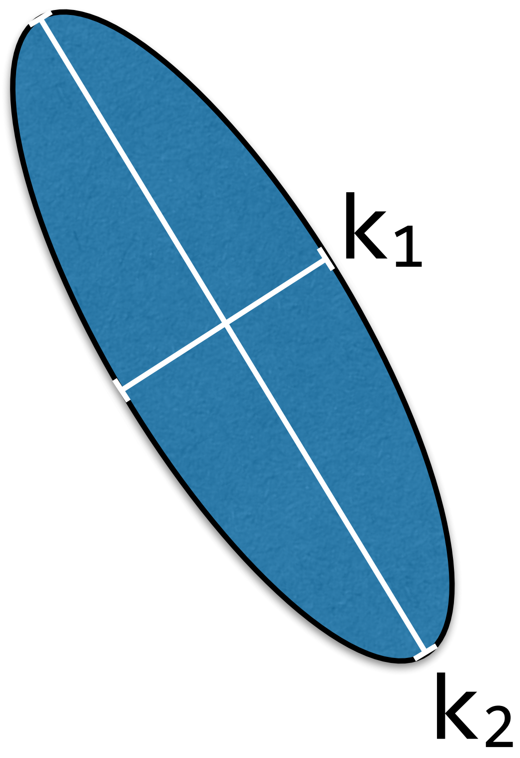
REGRESSION STABILITY



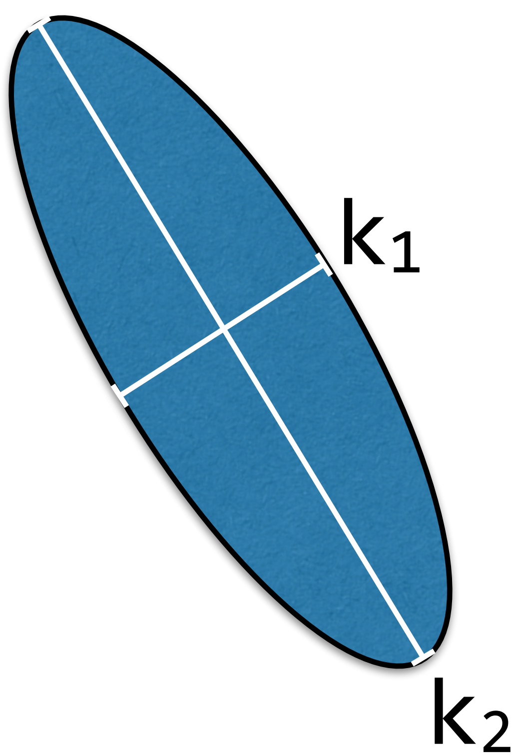
REGRESSION STABILITY

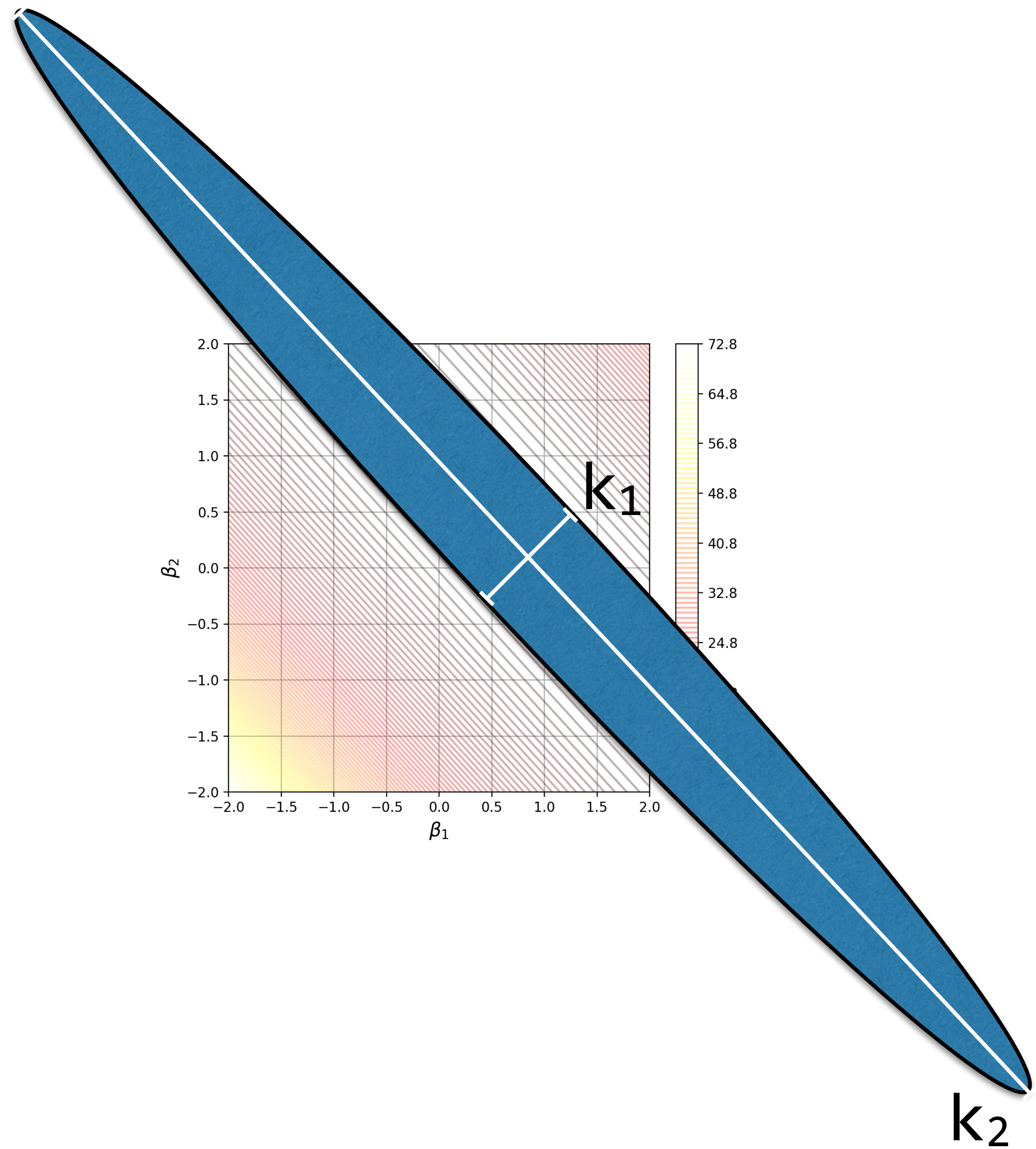


REGRESSION STABILITY



REGRESSION STABILITY





REGRESSION STABILITY

* the dimensions of the error ellipse, k_1 and k_2, are related to the **singular values** returned by np.linalg.lstsq!

* if the singular values (ordered from largest to smallest) are s_1, s_2, etc.,

* then $k_1 \propto s_1^{-1}$, $k_2 \propto s_2^{-1}$, etc.

REGRESSION STABILITY

* so it’s easy to detect when a regression is unstable: look for tiny singular values! (example)

* (or at least, tiny relative to the largest singular value)

REGRESSION STABILITY

* but what do you do if the regression is unstable?

* how could you possibly solve this problem?

END