Least Angle Regression on Diabetes Dataset

Amit Kumar

Machine Learning and Computing
Department of Mathematics
Indian Institute of Space Science and Technology, Trivandrum

July 15, 2021

Overview

1. Analysis of Dataset
2. LARS

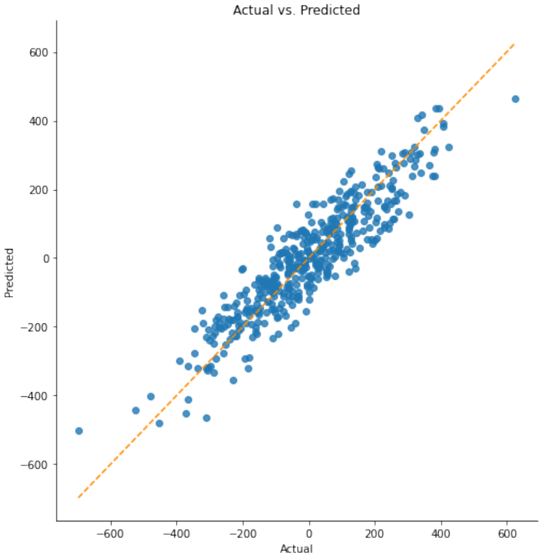
Dataset

- Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of $n = 442$ diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.
- Number of Attributes: First 10 columns are numeric predictive values
- Attribute Information:
  - Age
  - Sex
  - Body mass index
  - Average blood pressure
  - S1 (serum measurement 1)
  - S2 (serum measurement 2)
  - S3 (serum measurement 3)
  - S4 (serum measurement 4)
  - S5 (serum measurement 5)
  - S6 (serum measurement 6)
- Target: Column 11 is a quantitative measure of disease progression one year after baseline

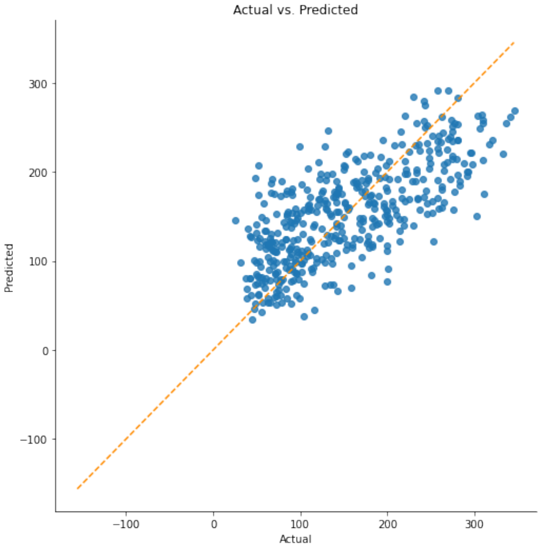
Linearity

- This assumes that there is a linear relationship between the predictors (e.g. independent variables or features) and the response variable (e.g. dependent variable or label). This also assumes that the predictors are additive.
- There may not just be a linear relationship among the data.
- If there doesn't exist linear relationship then the predictions will be extremely inaccurate because our model is underfitting. This is a serious violation that should not be ignored.

Linearity(Sample Data)



Linearity(Diabetes Data)



Normality of the Error Terms

- This assumes that the error terms of the model are normally distributed.
- A violation of this assumption could cause issues with either shrinking or inflating our confidence intervals.
- There are a variety of ways to do so, but we'll look at both a histogram and the p-value from the Anderson-Darling test for normality.

Anderson-Darling test

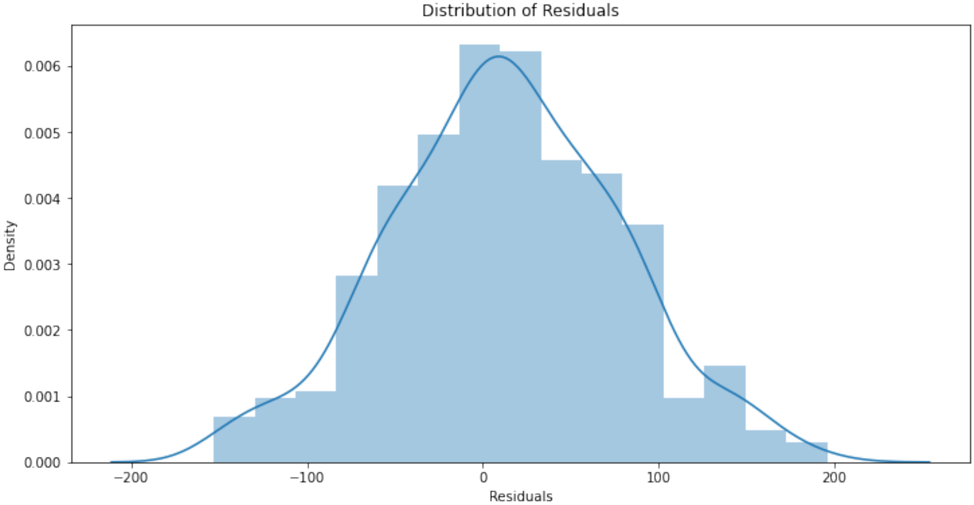
- The hypotheses for the Anderson-Darling test are:
  H0: The data comes from a normal distribution.
  H1: The data does not come from a normal distribution.

$$AD = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln F(X_i) + \ln(1 - F(X_{n-i+1}))]$$

- Where: n = the sample size, F(x) = CDF for the normal distribution, i = the ith sample, calculated when the data is sorted in ascending order

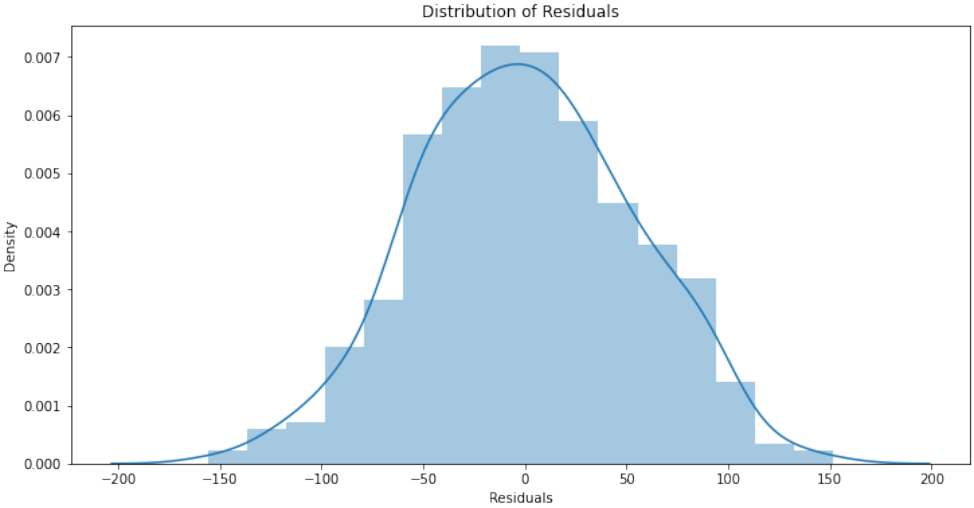
Anderson-Darling test(Sample Data)

- p-value from the test - 0.7835494346512862



Anderson-Darling test(Diabetes Data)

- p-value from the test - 0.43265797239371145



No Multicollinearity among Predictors

- This assumes that the predictors used in the regression are not correlated with each other.
- Multicollinearity causes issues with the interpretation of the coefficients. Specifically, we can interpret a coefficient as "an increase of 1 in this predictor results in a change of (coefficient) in the response variable, holding all other predictors constant."
- This becomes problematic when multicollinearity is present because we can't hold correlated predictors constant.

Variance inflation factor on Diabetes Data

- age: 1.2173065764321338
  sex: 1.2780725459826972
  bmi: 1.5094458375317008
  bp: 1.4594285821794586
  s1: 59.20378568651294
  s2: 39.19437938862489
  s3: 15.402352175616604
  s4: 8.89098622497626
  s5: 10.076221589049254
  s6: 1.4846225872940022
- 4 cases of possible multicollinearity
  0 cases of definite multicollinearity

Homoscedasticity

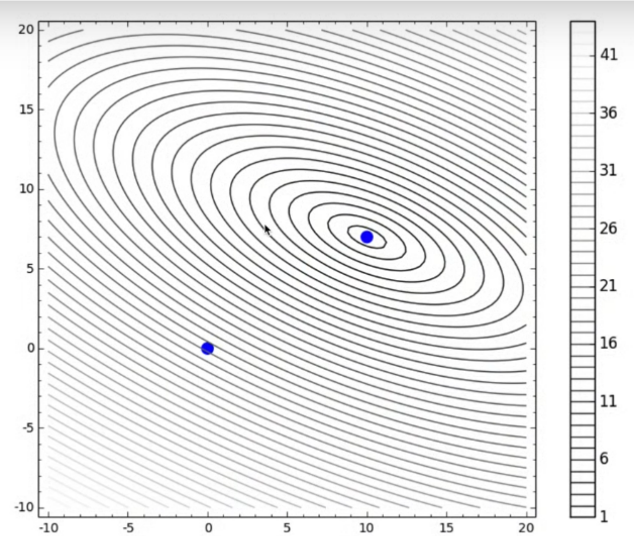
- This assumes homoscedasticity, which is the same variance within our error terms.
- Heteroscedasticity, the violation of homoscedasticity, occurs when we don't have an even variance across the error terms.
- It happens when our model may be giving too much weight to a subset of the data, particularly where the error variance was the largest.
- The confidence intervals will be either too wide or too narrow.
- Residuals should have relative constant variance.

Linear Data

- Initial Data:
  - Array of Response Y, shape n*1
  - Array of Predictor X, shape n*p
- Goal:
  - Find a model M, shape p*1 to write $Y \approx XM$
- Quality Measurement
  - Residual R(M)=Y - XM, shape n*1
  - Residual Sum of Square RSS(M)=R^TR

Best Linear fit

- Gauss-Markov: M_f is the unique unbiased minimizer of RSS(M)



Best Linear fit

- Best fit M_f should minimize error
  $0 = RSS_{\prime}(M_f\ ) =$ -2X^T(Y-XM_f) = -2X^TY $+$ 2X^TXM_f
- so the best fitting model M_f solves a linear equation

$$(X^T\mathrm{X})M_f\ = X^T\mathrm{Y}$$
$$M_f\ = (\mathrm{X}^T\mathrm{X})^{-1}X^T\mathrm{Y}$$

Best Linear fit ?

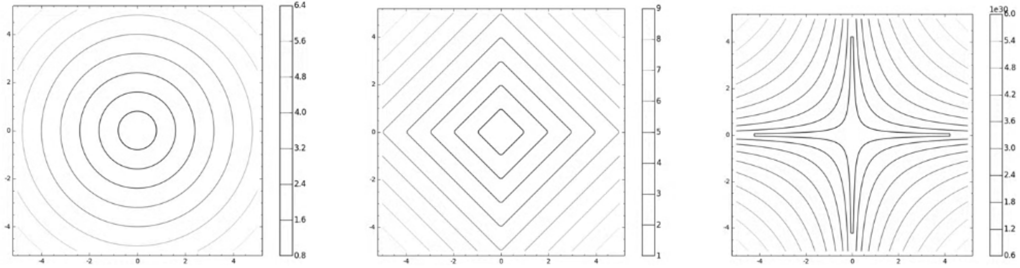
- $Y \approx 25.0012x1 + 0.0006x2 + 8.992x3 + ... - 387.345x1000$
  $Y \approx 22.006x1 + 8.7532x3 - 383.345x1000$
  $Y \approx 21.0012x1 - 360.345x1000$
  $Y \approx 0$

Practically we must balance accuracy against simplicity.

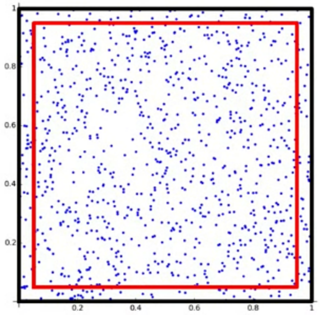
Best Linear fit ? New Goal

- Minimize RSS(M) among all M of a given complexity
- How to measure complexity? Need a norm

$\|v\|_{L_2}$	$\|v\|_{L_1}$	$\|v\|_{L_0}$
$\sqrt{\lvert v_1\rvert^2 + \cdots + \lvert v_n\rvert^2}$	$\lvert v_1\rvert + \cdots + \lvert v_n\rvert$	# of nonzero terms



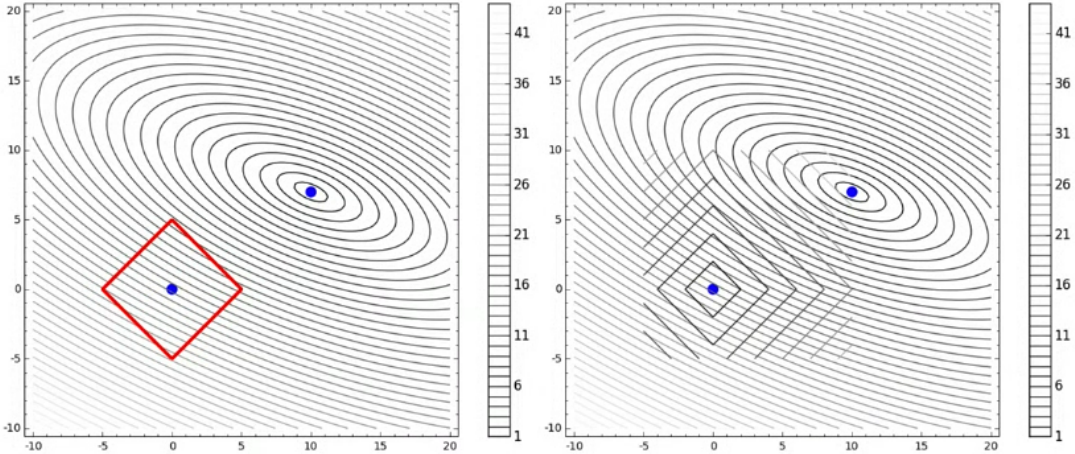
Curse of Dimensionality



- if n $= 2$, then $(0.90)^n = 0.81$
- if n $= 3$, then $(0.90)^n = 0.729$
- if n $= 10000$, then $(0.90)^n = 2.66 * 10^{-458}$
- Curse of Dimensionality means L_1 is almost L_0

LARS

- Fix a budget of $\|M\|_{L_1}$ then Minimize RSS(M)



- E_t(M)$=$ 1/2*RSS(M) $+$ t $\|M\|_{L_1}$

LARS Algorithm

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N-1, p)$ steps, we arrive at the full least-squares solution.

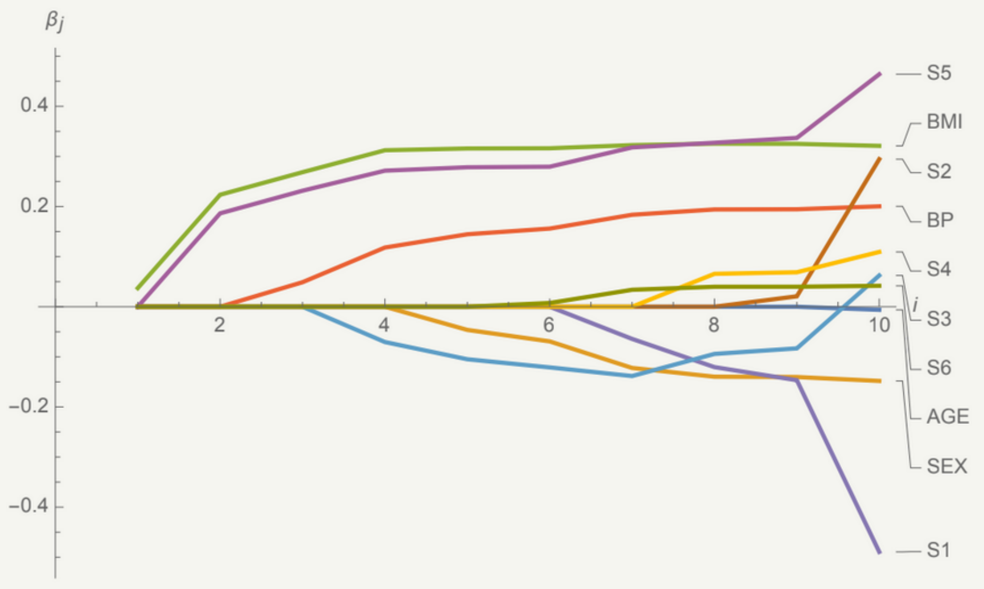
LARS Algorithm

- Assume standardized predictors in the model (mean 0 and unit variance)
    1. Start with no predictors in the model
    2. Find the predictor most correlated to the residual (equivalently, the variable making least angle with the residual)
    3. Keep moving in the direction of the most correlated predictor until another predictor becomes equally correlated with the residual.
    4. Move in a direction equiangular to both the predictors
    5. Continue until all the predictors are in the model

Modification of LARS from LASSO

- If a non-zero coefficient hits zero, drop its variable from the active set of variables and recompute the current joint least squares direction

Result



References

1. Efron, Bradley, et al. "Least angle regression." The Annals of statistics32.2 (2004): 407-499.
2. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2008.

The End