Multicollinearity

STA9750

Introduction to multicollinearity

If the predictors that are in the model are highly correlated, we can run into problems, because...

- The standard error of the coefficients increases
- Estimation becomes “unstable”

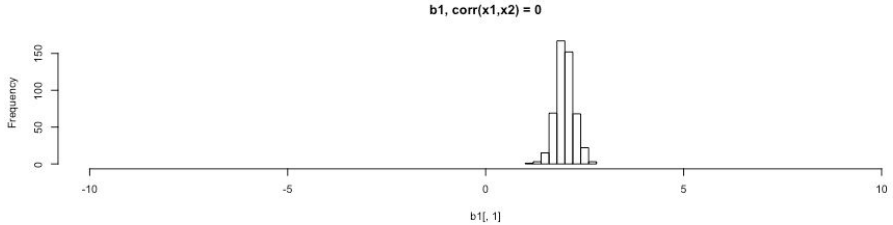
Example

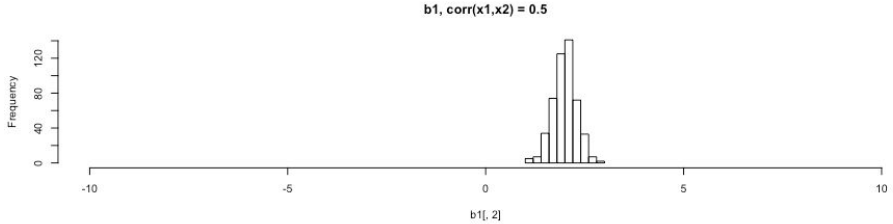
Model with 2 variables, x1 and x2, sample size n = 20

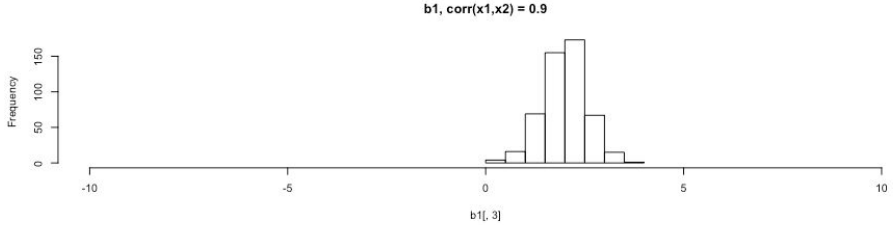
The truth is $y = 1 + 2\,x1 + 2\,x2 + \varepsilon$, with $\varepsilon \sim N(0,1)$

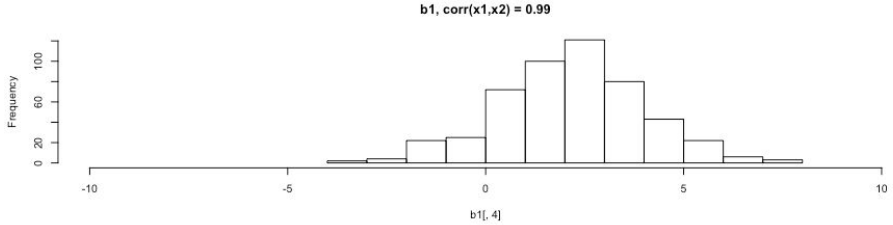
- Case 1: Corr(x1, x2) = 0
- Case 2: Corr(x1, x2) = 0.5
- Case 3: Corr(x1, x2) = 0.9
- Case 4: Corr(x1, x2) = 0.99

Simulate 1000 datasets and find estimated coefficients









Corr(x1,x2)	Var(b1)
0	0.06
0.5	0.09
0.9	0.31
0.99	3.27

Variance inflation factors (VIFs)

Regress covariate on the rest of covariates, find R squared

If R squared is close to 1, the variable is nearly linearly dependent on the other variables

$$VIF = 1/(1-R^2)$$

In PROC REG, add “/ VIF” to your model statement.

LASSO

Whereas the classical regression estimator solves the least squares problem

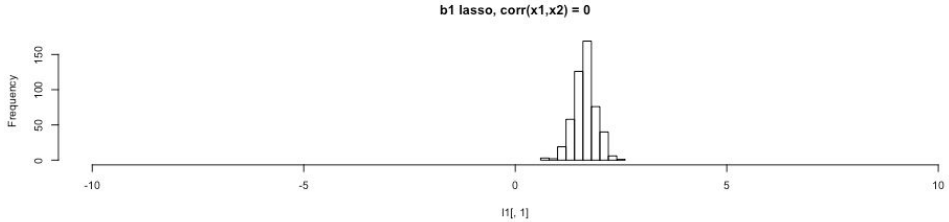
$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

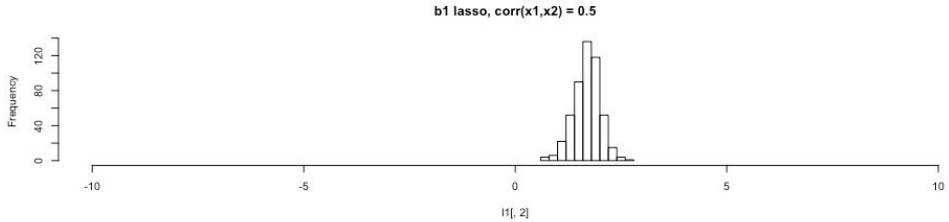
the lasso estimator solves the least squares problem by placing an ℓ_1 penalty on the parameters:

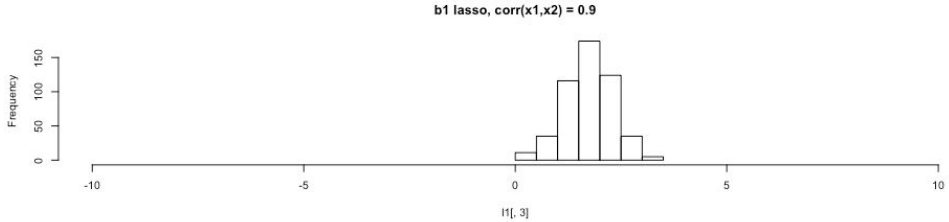
$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2$$

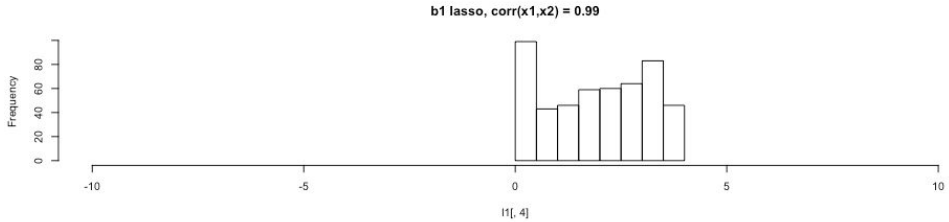
$$\text{subject to } \sum_{j=1}^{p} |\beta_j| \leq t$$

Source: https://analytics.ncsu.edu/sesug/2016/SD-253_Final_PDF.pdf









Corr(x1,x2)	Var(b1 lasso)
0	0.06
0.5	0.09
0.9	0.28
0.99	1.53

Splitting data into training and test

If you have a dataset that’s big enough, you can use a fraction of the data to do model building

And then, use the other fraction to test out and compare your models

K-fold cross-validation

Sometimes, you can’t afford to do that

K-fold cross-validation:

1. Split the data into K random subsets that have the same # of observations
2. Exclude one of the subsets, fit the model with the remaining (K-1) subsets
3. Fit your model and predict the observations you left out.
4. Find the predictive error (e.g. the sum of the mean squared errors)
5. Do steps 2 through 4 for all of the K subsets.
6. Average the K estimations of predictive errors

K-fold cross-validation

You can do k-fold cross validation for any kind of model, not just normal linear regression models

You can compare models with transformed outcomes (y) if you compute the predictive error in the original scale

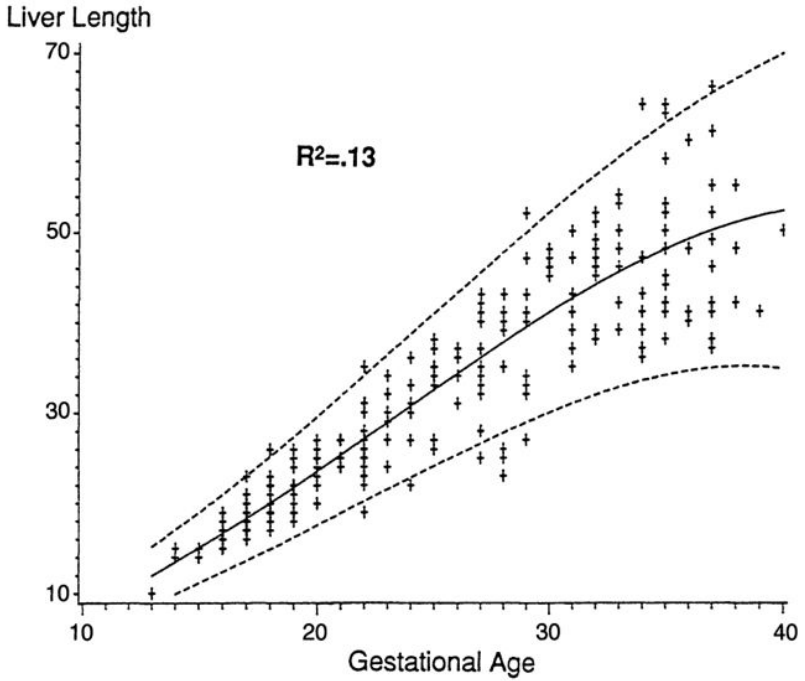


Figure 1. Fitted Model Based on $Y_1^* = Y/x^{3/2}$.

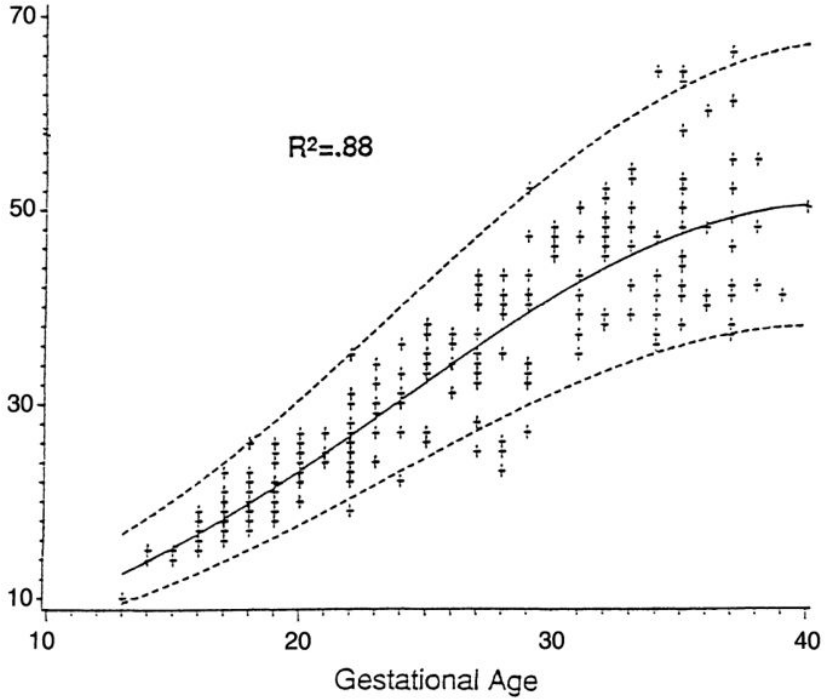


Figure 2. Fitted Model Based on $Y_2^* = \log Y$.

We can deal with situations like these with cross-validation