Predictive Analytics
Regression and Classification
Lecture 3 : Part 1

Sourish Das

Chennai Mathematical Institute

Aug-Nov, 2020

What is multicollinearity?

- Consider the standard linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$ and $n > p$

- This implies $\mathbf{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$

- The least square estimator of β is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

- The sampling distribution of $\hat{\boldsymbol{\beta}}$ is

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$$

What is multicollinearity?

- If correlation between two predictors of $\mathbf{X}$ is 1, that means one column is exactly dependent on other, that will result $det(\mathbf{X}^T\mathbf{X}) = 0$
- Heance $\mathbf{X}^T\mathbf{X}$ will not be invertible, (because $(\mathbf{X}^T\mathbf{X})^{-1} = \frac{Adj(\mathbf{X}^T\mathbf{X})}{det(\mathbf{X}^T\mathbf{X})}$
- In such case unique solution does not exists.

Why multicollinearity is a problem?

- If correlation between two predictors of $\mathbf{X}$ is nearly 1 or -1, **but not exactly 1**.
- For example $cor(X_i, X_j) = 0.99$ - what happens then?
- $det(\mathbf{X}^T\mathbf{X}) = \delta > 0$, where δ is a very small value.
- $\mathbf{X}^T\mathbf{X}$ is invertible - but every element of $(\mathbf{X}^T\mathbf{X})^{-1}$ **will be very large**.
- Unique solution $\hat{\beta}$ exists but $Cov(\hat{\beta}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ will be extremely large - so standard error will be very large.

  Hence valid statistical inference cannot be implemented.

Correlated Predictors

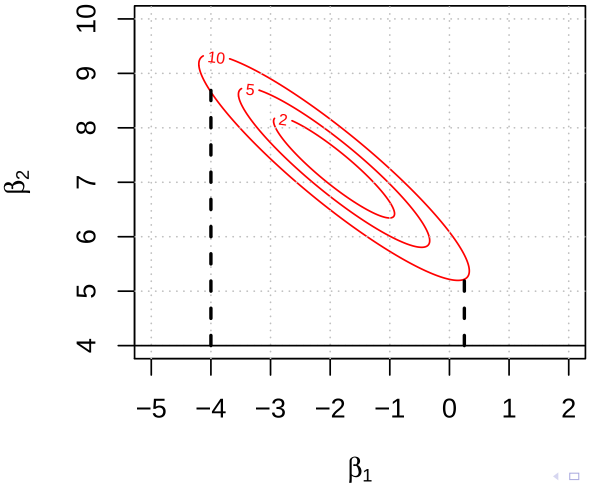
- We consider simple no-intercept model:

  $$\texttt{mpg} = \beta_1 \texttt{wt} + \beta_2 \texttt{drat} + \epsilon$$

- $\rho(\texttt{wt},\texttt{drat}) = -0.71$

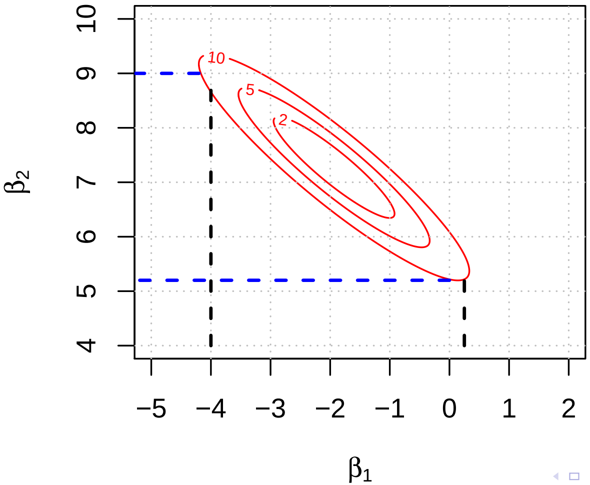
Sampling distribution for β_0 and β_1

OLS Estimator induces $\rho(\hat{\beta}_1, \hat{\beta}_2) = -0.92$



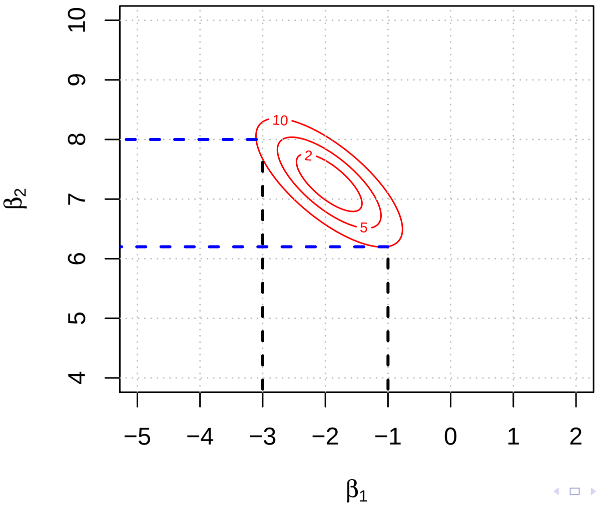
Sampling distribution for β_0 and β_1

OLS Estimator induces $\rho(\hat{\beta}_1, \hat{\beta}_2) = -0.92$



Sampling distribution for β_0 and β_1

Ridge Estimator induces $\rho(\hat{\beta}_1, \hat{\beta}_2) = -0.73$



Identify multicollinearity

- variance inflation factor (VIF) is an index which indicates how much a feature is contributing towards the multicollinearity problem
- Analyze the magnitude of multicollinearity by considering the size of the VIF($\hat{\beta}_i$) A rule of thumb is that if VIF($\hat{\beta}_i$) $>$10 then multicollinearity is high.
- A cutoff of 5 is also commonly used.

Variance Inflation Factor

- Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon.$$

- The standard error of $\hat{\beta}_j$ is

$$se(\hat{\beta}_j) = \sqrt{s^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$$

- It turns out that variance of $\hat{\beta}_j$ can be expressed as

$$\mathbb{V}ar(\hat{\beta}_j) = \frac{s^2}{(n-1)\mathbb{V}ar(X_j)} \cdot \frac{1}{1-R_j^2},$$

where R_j^2 is the multiple R^2 of X_j on $\{X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_p\}$, i.e.,

$$X_j = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_{j-1}X_{j-1} + \gamma_{j+1}X_{j+1} + \cdots + \gamma_p X_p + \epsilon$$

Variance Inflation Factor

- Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon.$$

- The standard error of $\hat{\beta}_j$ is

$$se(\hat{\beta}_j) = \sqrt{s^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj}}$$

- It turns out that variance of $\hat{\beta}_j$ can be expressed as

$$\mathbb{V}ar(\hat{\beta}_j) = \frac{s^2}{(n-1)\mathbb{V}ar(X_j)} \cdot \frac{1}{1-R_j^2},$$

- The term $\frac{1}{1-R_j^2}$ is known as the VIF of j^{th} predictor.

Implementation

- In R, the function `vif` in `car` package implemente the variance inflation factor.
- In `Python` the function `variance_inflation_factor` in `statmodels` can be used to identify the multicollinearity.

In the next part..

- We will discuss the issues of ill-posed problems...
- The problem of **multicollinearity** is a special case of a class of problems called ill-posed problems.

Thank You

sourish@cmi.ac.in