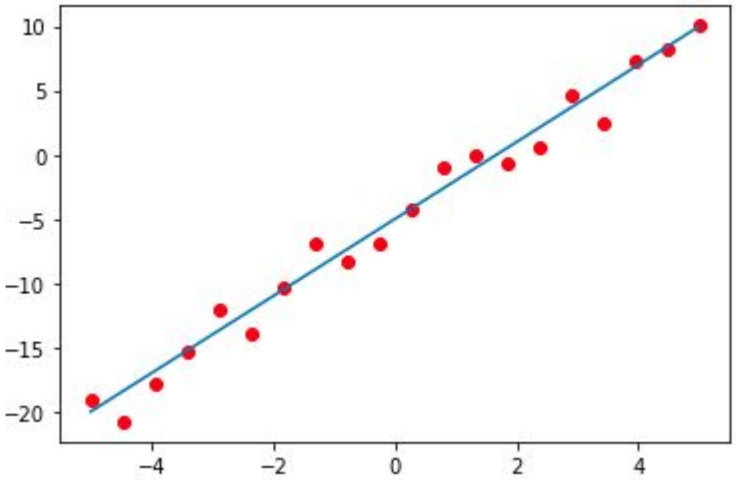
Multicollinearity; Diagnostics and Modification

GR 5205 / GU 4205
Section 3

Columbia University
Xiaofei Shi

Model assumptions



Why it is a problem...

- $\widehat{\beta} = (x^\top x)^{-1} x^\top Y$ problematic if not invertible.
- $\mathrm{Var}[\widehat{\beta}] = \sigma^2 (x^\top x)^{-1}$ going to blow up if close to singular.
- Collinearity v.s. Multicollinearity:

  - collinearity: lying in the same straight line

  - multicollinearity: one predictor variable in a multiple regression model can be linearly

  predicted from the others with a substantial degree of accuracy.

How to identify

There are several equivalent conditions for any square matrix $\mathbf{U}$ to be singular or non-invertible:

- The determinant $\det \mathbf{U}$ (or $|\mathbf{U}|$) is 0.
- At least one eigenvalue of u is 0. (This is because the determinant of a matrix is the product of its eigenvalues.)
- $\mathbf{U}$ is **rank deficient**, meaning that one or more of its columns (or rows) is equal to a linear combination of the other rows.

Since we're not concerned with any old square matrix, but specifically with $\mathbf{X}^T\mathbf{X}$, we have an additional equivalent condition:

- $\mathbf{X}$ is **column-rank** deficient, meaning one or more of its columns is equal to a linear combination of the others.

Pre-processing of predictors

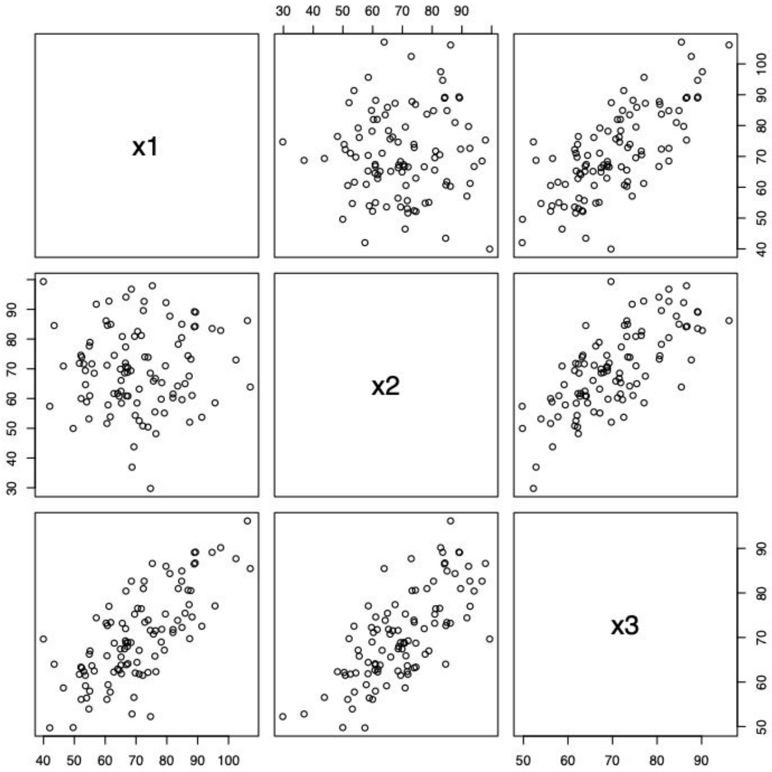
- Always normalize X first!
- Not all predictors are actually contributing information, a natural way of dealing with collinearity is to drop some variables from the model.
- Plotting pairs of predictors to see if there are collinearity: see if any of them fall on a straight line, or close to one; best way to detect when
- But multicollinearity is hard to detect....

Why multicollinearity is hard to see

$$X_1, X_2 \; i.i.d. \sim \mathcal{N}(0, \sigma^2)$$

$$X_3 = \frac{1}{2}(X_1 + X_2)$$

But we cannot really tell from the pair plots



Then what to do...

- $\mathbf{X}^T\mathbf{X}$ is special: symmetric and positive semidefinite

  - hence is diagonalizable

  - singular value decomposition (SVD) to find the eigenvalues, to check if 0 is contained

  - SVD also tells you how different your

- What about robustness?

Ridge Regression

- Model:
- Loss function:
- Estimator:
- is always invertible (why)?
- R package: MASS package lm.ridge; ridge package linearRidge

  Python package: sklearn.linear_model.Ridge

High-dim Regression

- High dimensional problem: $p>n$

  Big data: $n>>1$, $p>>1$

- General strategy: penalizing estimates or dimension reduction
- Algorithm for high-dim regression:

  - Forward stepwise regression

  - Ridge regression

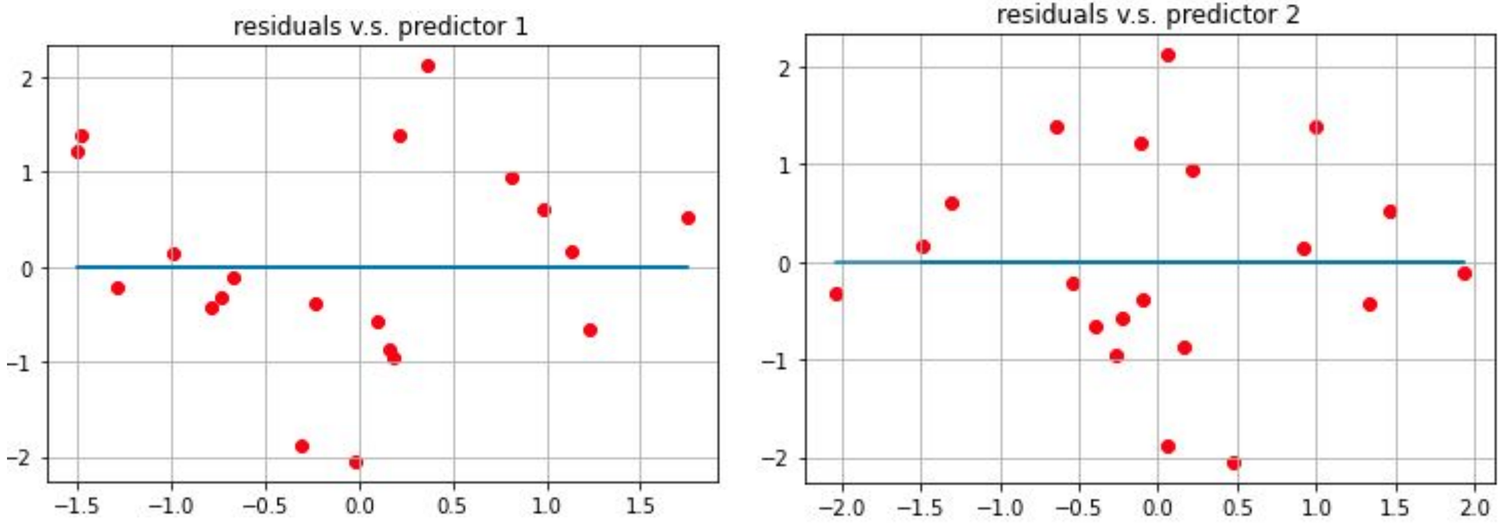
  - Lasso

Summary of Properties of Residuals

$$e = y - \hat{y} = y - xb$$

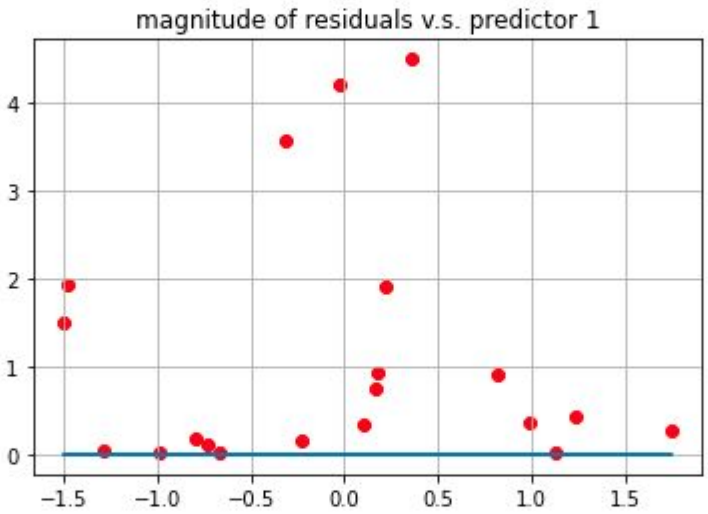
1. The residuals have mean 0, and they sum up to 0
2. The residuals should show a constant variance, unchanging with x
3. The residuals can’t be completely uncorrelated with each other, but the correlation should be extremely weak, and grow negligible as $n \to \infty$.
4. If the noise is Gaussian, the residuals should also be Gaussian.

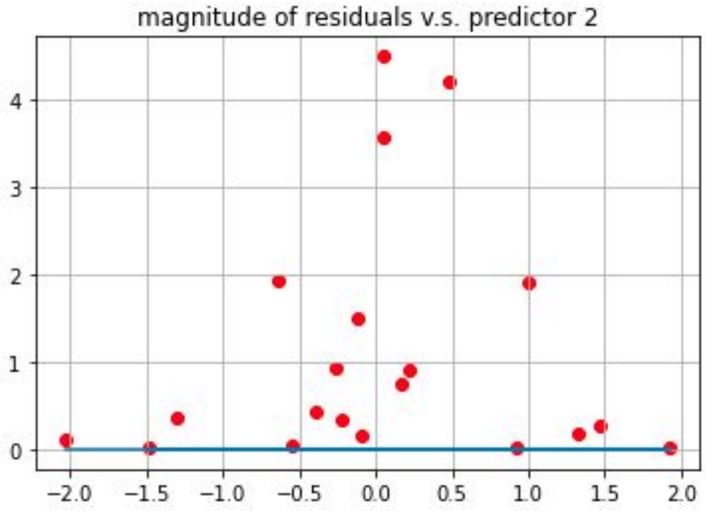
1. residuals v.s. predictors



evenly distributed around line 0 suggests residuals have mean 0

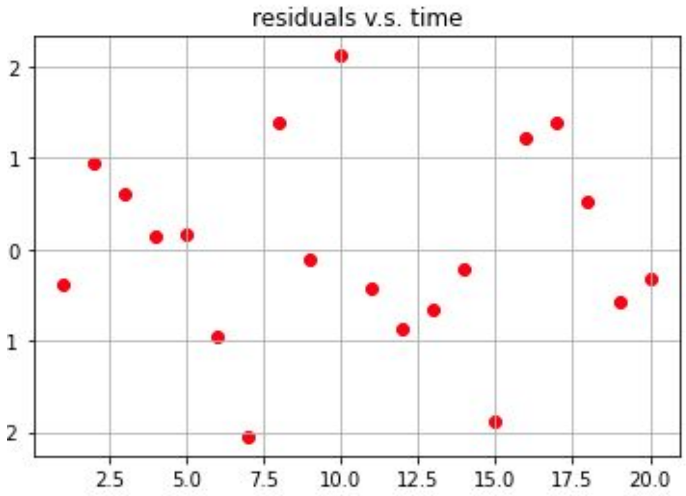
2. magnitude of residuals v.s. predictors

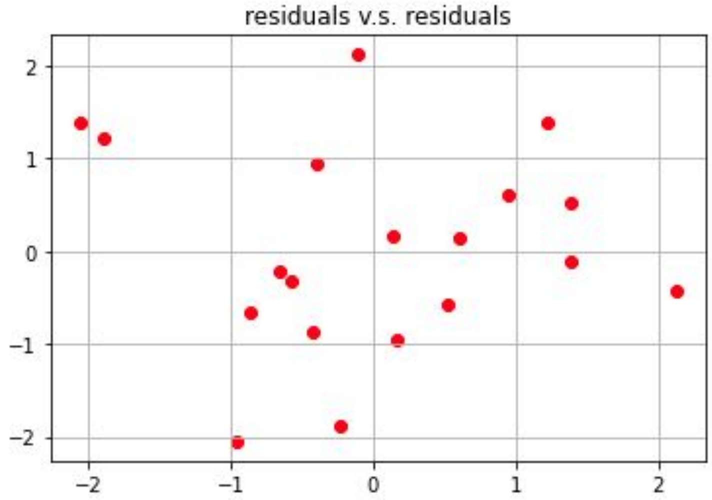




similar magnitude across different predictor values suggests constant variance

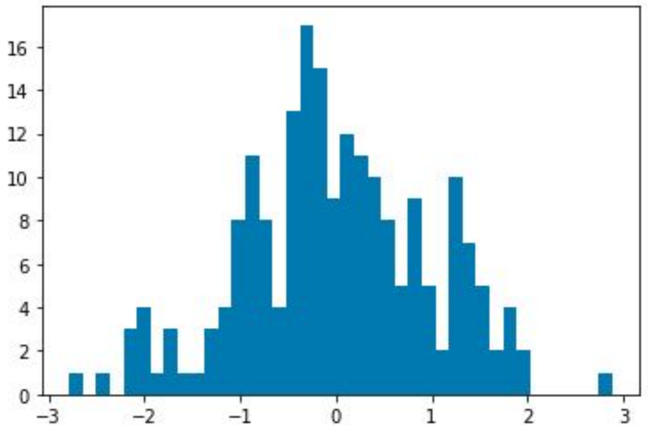
3. residuals v.s. residuals/coordinates

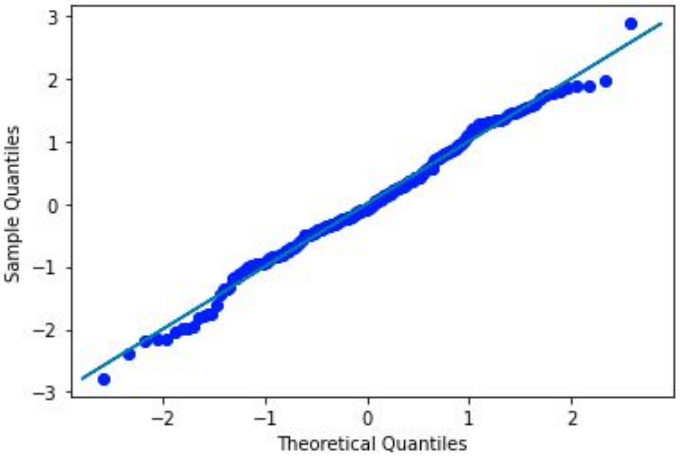




no trend at all suggests non-correlated residuals

4. distribution of residuals: histogram of residuals and Q-Q plots





similar to normal density/ Q-Q plot lies on x=y suggests the Gaussian residuals hold.

More about part 4:

- Q-Q plots for other distributions: Cauchy, etc
- Q-Q plots for two data distributions
- P-P plots
- Formal tests: Chi-square test for histogram, K-S test for Q-Q plot

Generalization

- cross-validation
- training set and testing set

Nonlinearity of Y v.s. X

- Transformations
- Nonlinear least square: WLS
- Smoothing
- Generalized linear models

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} y_i I_{[x-h,x+h]}(x_i)}{\sum_{j=1}^{n} I_{[x-h,x+h]}(x_j)}$$